Science of the Total Environment 672 (2019) 438–455

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Review

"-Omics" workflow for paleolimnological and geological archives: A review

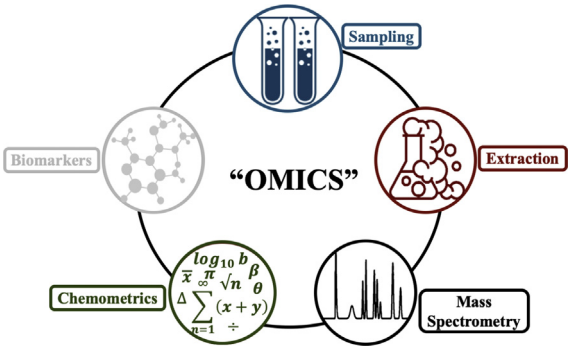
Madison Bell, Jules M. Blais *

Laboratory for the Analysis of Natural and Synthetic Environmental Toxicants, Department of Biology, University of Ottawa, Ottawa, ON K1N 6N5, Canada

HIGHLIGHTS

- "-Omics" workflow with examples relevant to sediment, soil and oil samples.
- Highlight practical considerations for each step in the "-omics" workflow.
- Establish quality control and standard practices for environmental "-omics".
- Focus on production of reliable and statistically powerful "-omics" data.
- Overview of common univariate and multivariate statistical analyses.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 1 November 2018
Received in revised form 29 March 2019
Accepted 30 March 2019
Available online 2 April 2019

Editor: Yolanda Picó

Keywords:
Proxies
Biomarkers
Untargeted
Sedimentomics
Petroleomics
Chemometrics

ABSTRACT

"-Omics" is a powerful screening method with applications in molecular biology, toxicology, wildlife biology, natural product discovery, and many other fields. Genomics, proteomics, metabolomics, and lipidomics are common examples included under the "-omics" umbrella. This screening method uses combinations of untargeted, semi-targeted, and targeted analyses paired with data mining to facilitate researchers' understanding of the genome, proteins, and small organic molecules in biological systems. Recently, however, the use of "-omics" has expanded into the fields of geology, specifically petrology, and paleolimnology. Specifically, untargeted analyses stand to transform these fields as petroleomics, and sediment-"omics" become more prevalent. "-Omics" facilitates the visualization of small molecule profiles from environmental matrices (i.e. oil and sediment). Small molecule profiles can provide improved understanding of small molecules distributions throughout the environment, and how those compositions can change depending on conditions (i.e. climate change, weathering, etc.). "-Omics" also facilities discovery of next-generation biomarkers that can be used for oil source identification and as proxies for reconstructing past environmental changes. Untargeted analyses paired with data mining and multivariate statistical analyses represents a powerful suite of tools for hypothesis generation, and new method development

Abbreviations: AMDIS, Automated Mass Spectral Deconvolution and Identification System; ANOVA, analysis of variance; APCI, atmospheric pressure chemical ionization; APPI, Atmospheric Pressure Ionization; CCS, Collisional Cross Section; COLMAR, Complex Mixture Analysis by NMR; DBE, double bond equivalents; DIA, Data Independent Acquisition; EI, electron impact; ESI, electrospray ionization; FT-ICR, Fourier transform ion cyclotron resonance; GCMS, Gas Chromatography Mass Spectrometry; GCxGC, two-dimensional gas chromatography; GMD, Golm metabolome database; HCA, hierarchical clustering; HMDB, Human Metabolome Database; HRMS, High Resolution Mass Spectrometry; HSD, Honestly Significant Difference; IMMS, ion mobility mass spectrometry; IR, infrared spectroscopy; LCxLC, two-dimensional liquid chromatography; LCMS, Liquid Chromatography Mass Spectrometry; LRMS, Low Resolution Mass Spectrometry; LSD, Least Significant Difference; MCCV, Monte Carlo cross-validation; MMCD, Madison Metabolomics Consortium Database; NIST, National Institute of Standards and Technology; NMR, Nuclear Magnetic Resonance; OPLS, Orthogonal Partial Least Squares or Projection to Latent Structures; OM, organic matter; PCA, principal component analysis; PLS, partial least squares or projection to latent structures; QC, quality control; MS^n, tandem mass spectrometry; RF, random forest; ROC, receiver operating characteristic; SOM, self-organizing maps; sPLS, Sparse Partial Least Squares or Projection to Latent Structures; SVM, support vector machines; TOF, time of flight; VIP, variables of importance.

* Corresponding author.
E-mail address: jules.blais@uOttawa.ca (J.M. Blais).

https://doi.org/10.1016/j.scitotenv.2019.03.477
0048-9697/

for environmental reconstructions. Here we present an introduction to "-omics" methodology, technical terms, and examples of applications to paleolimnology and petrology. The purpose of this review is to highlight the important considerations at each step in the "-omics" workflow to produce high quality and statistically powerful data for petrological and paleolimnological applications.

Contents

1. Introduction

1.1. "-Omics"

The field of "-omics" research is relatively young, starting in the 80's with genomics (Robertson, 2005; Yadav, 2007). The origin of the term "-omics" may originate from the Sanskrit "*OM*" meaning completeness (Yadav, 2007). Thus, when combined with the targets under study, it implies "all genes" for genomics, or "all lipids" for lipidomics. Currently, the field of "-omics" is extensive and includes: genomics, transcriptomics, proteomics, metabolomics, lipidomics, metallomics, and environment-omics applied to environmental samples. The focus of this review is on environmental "-omics", and how to apply "-omics" techniques to environmental matrices like sediment, soil, peat, oil, etc. "-Omics" is a holistic approach to studying biological systems and refers to an interdisciplinary methodology that incorporates analytical chemistry, statistical modelling and multivariate analyses, and knowledge of the systems under survey.

The advantage of using "-omics" is the generation of novel hypotheses. "-Omics" investigations use untargeted analyses to generate a "broad compositional" or "global" view of the system. This untargeted global survey of the targeted molecules (i.e. commonly DNA, proteins, or small molecules) allows for the assessment of broad patterns based on compositional changes. Successive data mining of these patterns can pinpoint where these compositional changes are, which can lead to the identification of potential biomarkers. Biomarkers are organic molecules that are significantly different in either abundance or presence between the sampled groups. Thus, the utility of an untargeted "-omics" approach is to generate new hypotheses regarding these "global" patterns and potential biomarkers.

Semi-targeted and targeted analyses explore the new hypotheses to determine if they can stand up to rigorous scientific validation. For

instance, a targeted analysis of a potential biomarker would answer more specific questions like: Does this biomarker only appear in oil from this region? Or does this biomarker preserve over time in sediment samples? Ideally, follow-up targeted analyses validate the hypotheses generated with the untargeted "global" analysis. The "-omics" framework of untargeted "global" analyses followed by semi-targeted and targeted analysis is widely applicable to many fields.

1.2. Current applications of "-omics" in environmental sciences

Currently, the use of "-omics" in environmental fields is burgeoning. There are four main areas in which this methodology has been applied to sediment, soil, and oil matrices: (1) Untargeted screening for environmental contaminants; (2) Investigation of soil and sediment bacterial community metabolites and carbon cycling; (3) Untargeted screening of organic matter (OM) composition leading to novel biomarkers for paleolimnological applications; and (4) petroleomics or oil fingerprinting.

Semi-targeted and untargeted screening approaches have been used on sediment or soil samples for pollutants analysis. There are many studies documenting the screening of sediment for flame retardants (Gustavsson et al., 2018), sediment-bound and unbound xenobiotics (Kronimus and Schwarzbauer, 2007), and anthropogenic contaminants in wastewater, sludge and sediment (Grigoriadou and Schwarzbauer, 2011; Rostkowski et al., 2013; Veenaas and Haglund, 2017). Many of these untargeted methods are not considered "-omics" methods, because they do not incorporate discriminant statistical analyses into their studies and are not focussed on hypothesis generation, but rather the confirmation pollutants are present. However, current untargeted and semi-targeted pollutant screening methodologies paired with "-omics" statistical analyses (Section 4.3.3) may facilitate pollutant discovery and source identifications with the appropriate study design and extraction techniques to compensate for complex matrices and trace levels of pollutants.

Soil and sediment microbial communities and metabolites can also be investigated with "-omics". Beale et al. (2017) demonstrated a multi-"omics" approach combining genomics and sediment-"omics" to assess the impact of environmental toxicants on bacterial colonies. A later publication compared the effect of changes in nutrients, OM, light and pollutants on bacterial populations, and found precipitation had a major influence on bacterial community structure and linked OM composition to bacterial metabolic pathways (Beale et al., 2018). Swenson et al. (2018) also linked the impact of wetting events on soil microbial community structures with changes in water extractable metabolites. Furthermore, "-omics" can be paired with stable isotope labelling to investigate microbial carbon cycling (Swenson et al., 2015). They were able to differentiate the isotope-labelled microbially-derived metabolites from endogenous soil OM and determine the extent of microbial activity (Swenson et al., 2015).

Other OM components, aside from microbial metabolites and contaminants, have undergone "-omics". For instance, an untargeted approach was used to investigate changes in soil OM composition after fire-exposures and they were able to isolate the fire affected organic molecules in different particle size fractions (Jiménez-Morillo et al., 2018). Similar methods have also been applied to soil OM (Kujawinski, 2011; X.M. Li et al., 2018; Z. Li et al., 2018; Ward and Cory, 2015). Sedimentary OM have had environmental changes assessed in sediments using "-omics". Untargeted "-omics" led to the discovery of novel proxies for climate reconstructions. Farrés et al. (2015a) used an "-omics" method on marine sediment where they correlated long chain *n*-alkanes, long and short chain *n*-alkan-1-ols, alkenols, cholesterol, and squalene with down-core sea surface temperatures.

Lastly, "-omics" has been applied to oil. Petroleomics is useful for both identifying biomarkers specific to an oil deposit, and also for determining sources of oil spills (Wang et al., 2006). Thus far, untargeted "-omics" techniques have differentiated Brazilian crude oils (Kiepper et al., 2014; Laakia et al., 2017; Prata et al., 2016), Columbian oils (Silva et al., 2011), Alberta oil sands (Yang et al., 2011), Jianghan basin, and Nanyan basin oils (Zhang et al., 2015). It can also be applied to oil weathering processes. Hall et al. (2013) used two-dimensional gas chromatography (GCxGC) mass spectrometry to differentiate *Deepwater Horizon* oil post- and pre-weathering. They were able to identify the precursors of some of the oxygenated weathering products using partial least squares regression (Hall et al., 2013).

1.3. Purpose

Currently, "-omics" is widely applied in many fields, but we emphasize methodologies for small organic molecules (<1000 kDa) located in environmental matrices (i.e. soil, sediment, and oil). The purpose of this review is to provide an "-omics" workflow for paleolimnological and geological research. We will highlight important considerations at each step in the "-omics" workflow (Fig. 1) including sample preparation, chemical analysis, data mining and structure identification. We emphasize the quality control steps required to produce reliable and statistically powerful data. In addition, each section cites more specific

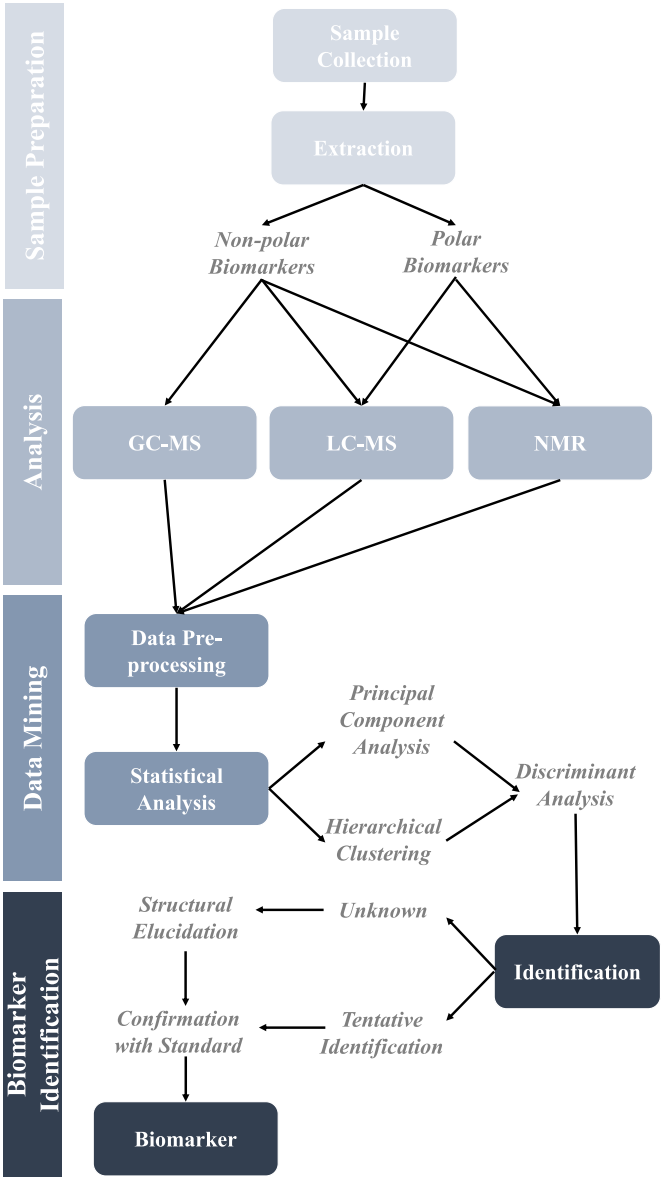


Fig. 1. A generalized workflow for small molecule untargeted "omics" strategies.

reviews for interested readers. The overarching objective is to elaborate on the potential of "-omics" research, and to establish practical protocols for future "-omics" studies on paleolimnological and geological samples.

2. Sample preparation

2.1. Experimental design

2.1.1. Sources of experimental variation

"-Omics" can be susceptible to variability at all stages of the workflow (Dudzik et al., 2018; Martins et al., 2018; Szczepańska et al., 2018). Good experimental design requires the consideration of all stages of the workflow (Fig. 1) before experimentation begins. This review emphasizes each stage of the workflow with the aim to reduce experimental variability.

There are five major types of variation for "-omics" in environmental matrices: (1) Sample variation – is due to samples' spatial relatedness (e.g. distances between sites or within cores), weathering, diagenesis, time between sample collections, environmental conditions during sample collection and sample collection techniques. (2) Pre-analytical variation – is caused by inconsistent or ill-considered sample extraction procedures, chemical modifications during extraction, pre-extraction sample storage and post-extraction degradation during storage. (3) Analytical variation – is caused by the presence of plasticizers, instrumental variation, batch effects, matrix effects causing ion suppression or enhancement and carryover effects from one sample to the next. (4) Post-analytical variation – includes data handling errors, false discoveries, mis-annotation of compounds, and data preprocessing errors. Lastly, (5) Stochastic variation – variation due to random chance. Untargeted "-omics" techniques are sensitive to many of these errors due to the nature of trying to "see" as much as possible in one sample, so experimental design is critical to minimize variation.

2.1.2. Untargeted, semi-targeted and targeted analytical strategies

There are three types of analysis strategies: (1) untargeted; (2) semi-targeted; and (3) targeted (Fig. 2). Untargeted strategies, also known as profiling, fingerprinting, or global assessments, are qualitative assessments of the molecular composition. The focus of an untargeted analysis is to see as many compounds as possible in an extract. Data mining identifies similarities and differences among the samples. Examples of relevant untargeted analyses include the work done by Farrés et al. (2015a, 2015b) on the discovery of climate biomarkers, and Alizadeh et al. (2017) who demonstrated a fractionated untargeted analysis of petroleum. This type of analysis is crucial for hypothesis generation and the discovery of potential biomarkers. Afterward, semi-targeted and targeted strategies validate the new hypotheses and biomarkers.

Semi-targeted strategies can be quantitative, or at least semi-quantitative. This strategy requires more sample preparation than untargeted approaches but has fewer analytical artifacts such as matrix effects, carryover, and coelution. In a semi-targeted analysis, the focus is on multiple known compounds, potentially from different compound classes. Examples of semi-targeted strategies include publications on the metabolomics profiles of polar bears (Morris et al., 2018) and screening organic contaminants (Bu et al., 2014; Gustavsson et al., 2018). The advantages of semi-targeted approaches are increased sensitivity, resolution, and quantification. The main disadvantage, when compared to an untargeted strategy, is the absence of novel compound discovery outside of the classes of interest.

Known molecular compounds undergo targeted analyses. Often there are multiple stages of sample clean-up, which will further reduce matrix effects, carryover, and coelution issues. Targeted analyses are quantitative with sensitive, precise, and accurate detection and quantification. These targeted studies validate biomarkers for future studies. The disadvantage of targeted analyses is that new biomarker discovery does not occur, and they require labour intensive extractions, which increases time to analysis. This review focuses on untargeted strategies, but there are detailed reviews for semi-targeted and targeted workflows specifically (Cajka and Fiehn, 2016; Gorrochategui et al., 2016).

2.1.3. Sample collection

Sample collection is an important consideration of experimental design for untargeted "-omics". A clear question is necessary before sampling. If the question is relevant to the molecular fingerprinting of different samples (Section 4.3.1), then contrast between the sampled groups does not need to be as defined (i.e. sampling a gradient of impact). If the question is relevant to biomarker discovery, then maximizing the differences between "Group 1" and "Group 2" (i.e. impacted versus unimpacted) during sample collection is necessary. High sampling contrast also optimizes the ability of discriminant statistical analyses (Section 4.3.3) to isolate potential biomarkers from the large amounts of data collected.

For sediment studies, there are additional considerations for sample collection. Sediment OM undergoes diagenesis at the sediment surface,

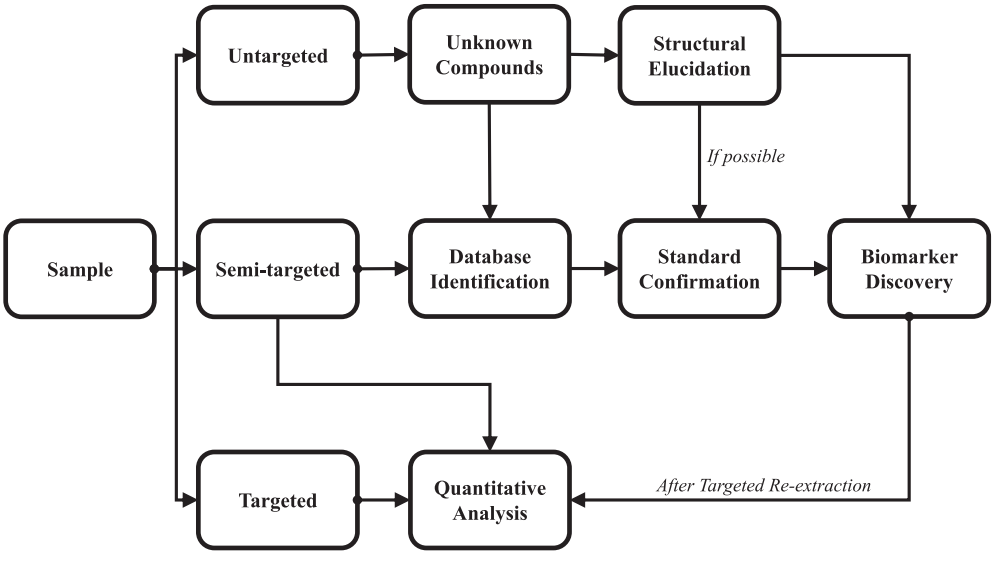


Fig. 2. Comparison workflow for targeted, semi-targeted, and untargeted strategies.

so the OM in the surface layer of a sediment core will not be the same as a layer within the anoxic layer after diagenesis (Meyers and Ishiwatari, 1993). There is also a temporal component as the sediment ages down-core; a single core can represent many years of sediment deposition. Thus, paleolimnologists must consider spatial sampling, but also temporal sampling as well. For example, if the research question addresses spatial changes, then temporal variation needs to be controlled by sampling layers in similar time frames. A proposed solution would be to use a top-bottom approach to account for diagenesis and environmental changes over time where a surface sediment sample and a down-core sediment sample are simultaneously analyzed. If the research question relates to temporal changes over time (i.e. oil weathering over some duration or OM changes during diagenesis) then the spatial variation needs to be minimized like in any time series or repeated measure experiments (Berk et al., 2011). Farrés et al. (2015a) provides an example of monitoring climatic signals over time, but limited in space, to correlate biomarkers with sea surface temperatures.

In addition, the number of samples requires careful consideration. It is prudent to perform a statistical power analysis before the start of the study. Greater sample sizes result in a reduced standard error and better precision in the results (Dudzik et al., 2018; Krzywinski and Altman, 2013). Of course, practical, budgetary, and resource constraints need consideration, but a statistical power analysis can help balance ideal sample size and prudence. The result should be a statistical power of 0.8, which is where, if the experiment is repeated multiple times, then there will be a significant difference between the two groups 80% of the time (Xia and Wishart, 2016). There are different tools and tests that can be used to perform power analysis for "-omics" studies and for more information see Xia and Wishart (2016), Blaise et al. (2016) and Dudzik et al. (2018).

2.2. Extractions

2.2.1. Non-polar extractions

The choice of extraction method is important, but it is always a compromise. No single extraction will extract all compounds. This section explores non-polar (i.e. lipid, organic contaminants, large terpenes) and polar (i.e. amino acids, small terpenes, carbohydrates) compound extraction methods. Table 1 contains examples of extraction methods for "-omics" on environmental matrices.

A non-polar extraction method may be more relevant for paleolimnological and geological biomarkers. Non-polar compounds preserve better in the sediment, especially saturated and aromatic compounds (Meyers and Ishiwatari, 1993) and oil samples largely contain very non-polar compounds. To demonstrate this, the majority of current biomarkers for paleolimnology can be seen in Fig. 3 in the context of their relative polarity and the solvent likely to extract them.

Currently, there is no review of ideal "-omics" solvent extractions for environmental matrices, but the most commonly used lipid extraction methods in biological matrices are liquid-liquid extractions derived from the Folch (Folch et al., 1956), Bligh-Dyer (Bligh and Dyer, 1959), and modified versions of both (Table 1; Alshehry et al., 2015; Cequier-Sánchez et al., 2008; S. Chen et al., 2013; Y. Chen et al., 2013; Löfgren et al., 2016; Matyash et al., 2008). In addition, dichloromethane or *n*-hexanes are the solvents of choice for oil, sewage, sludge, and sediment samples (Table 1). In a lipidomics study, Reis et al. (2013) determined solvent composition had little effect on predominant lipids, but was important for less abundant lipids. Sediment-"omics" and petroleomics studies may have the same solvent effects.

There are only a few examples of extraction methods for untargeted analyses in environmental matrices. One example used dichloromethane and fractionation into neutral and acidic groups to identify novel paleoclimatic biomarkers in marine sediments (Farrés et al., 2015a). Only a few studies have compared different extraction methods (Gregory et al., 2012; Mopper et al., 2007; Swenson et al., 2015; Tfaily et al., 2015, 2017; Warren, 2015). Tfaily et al. (2015, 2017) explored parallel and sequential extraction techniques and found the order of solvent extraction matters. For oil samples, Alizadeh et al. (2017) classified 27 different oils from the Southern Persian Gulf Basin and were able to separate the oils into two distinct families after extraction with *n*-hexanes then fractionation. Nizio et al. (2012) provided additional extraction strategies for petroleum, and many of the methods presented there may also be relevant to sediment samples.

Other methods include solid phase extraction (SPE) and solid phase microextraction (SPME; Cajka and Fiehn, 2016; Jurowski et al., 2017). Semi-targeted and targeted approaches typically use SPE and SPME for post-extraction clean-up and/or additional fractionation. There are advantages to using these methods such as reduced ion suppression, removal of interferents, and less solvent. SPME also excels in situations where the organic content is low or the sample size is small (Jurowski et al., 2017; Zhao et al., 2014). The major disadvantage is SPE and SPME can result in the loss of potentially discriminant compounds through retention and increased sample handling. Fractionation is not typical for traditional untargeted "-omics" methods; however, oil and sediment matrices are very complex, and fractionation may be required in order to avoid damaging sensitive instrumentation. Rostkowski et al. (2013) demonstrated the use of SPE cartridges and dichloromethane to assess different matrices like sewage, wastewater, sediment and sludge.

2.2.2. Polar extractions

Polar extraction methods extract compounds such as amino acids, flavonoids, hormones, and some amphoteric lipids, etc. (Fig. 3). Many metabolomics experiments are also designed to minimize protein and lipid content (Cajka and Fiehn, 2016; Raterink et al., 2014). However, polar extraction methods can also be applied to environmental matrices, especially if there is interest in linking microbial activity to metabolites excreted into the environment. Water and/or methanol extractable fractions often contain many compounds relevant to microbial communities. Example polar extractions can be found for sediment (Beale et al., 2017, 2018) and soil (Swenson et al., 2015, 2018; Warren, 2015). Also, Rostkowski et al. (2013) used a mid-polar approach (methanol:acetonitrile) on sewage, wastewater, sediment and sludge matrices to find organic contaminants.

2.2.3. Extraction quality control

Quality control (QC) is an important aspect of any untargeted analysis to reduce variability and noise. In order to reduce experimental variability, there has been a unified movement into established good experimental and reporting practices for "-omics", and there are a number of excellent reviews (Dudzik et al., 2018; Dunn et al., 2012). Every level of the workflow requires QCs. At the extraction level it is useful to account for pre-analytical variability due to extraction technique and operator error. Extraction QC includes:

- **Standard Operating Procedures** (**SOPs**) ensures future extractions follow the same protocols as closely as possible. SOPs should contain the information suggested for the minimum reporting standards (Sumner et al., 2007).
- **Environmental replicates** are parallel measurements from the same area but different locations (i.e. from the same oil field). These replicates account for random and spatial variation (Dudzik et al., 2018). Replicates should be in triplicates ($n = 3$), but if possible, more should be incorporated ($n = 5$; Sumner et al., 2007). Study failure can result from too few replicates when sample variance is high; however, these replicates may not always be possible. For instance, sediment samples from the same lake can vary due to sedimentation rates, and bathymetry.
- **Sample replicates** are from the same sample, but extracted in triplicate ($n = 3$) and are used to judge sample extraction consistency (Dudzik et al., 2018). When environmental replicates are not possible, then sample replicates are strongly encouraged to account for both extraction consistency and unknown variation.
- **Method blanks** are "blank samples" that undergo the same extraction protocol, but without sample (Dudzik et al., 2018). They account for

Table 1
Examples of extraction and analysis protocols for non-polar compounds and polar compounds.

Matrix	Molecular classes[a]	Extraction solvent[b]	Resuspension solvent	Analysis[c]	Reference
Non-polar extractions					
Wastewater, sewage, sediment	OC	DCM	Direct injection	GCxGC-TOF & GC-TOF	Rostkowski et al., 2013
Sediment	SH, CE, OH, SQ	DCM; fractionation	Toluene	GC–MS	Farrés et al., 2015a
Soil	OS, AA, P, SH, TN, LN, FA	Parallel H_2O, MeOH, ACN, hexane	Varied	ESI-FTICR/MS	Tfaily et al., 2015
Soil	OS, AA, P, SH, TN, LN, FA	Sequential H_2O, MeOH, ACN, hexane	Varied	ESI-FTICR/MS	Tfaily et al., 2017
Sewage	SH, OC, PAH, PL	4:1 *n*-hexane:DCM	Isooctane	GCxGC-TOF	Veenaas and Haglund, 2017
Feces	GP, FA, SL, GL, BA, CE, PL	1:4:3 HFIP:MTBE:H_2O (MTBE fraction)	65:30:5 ACN:IPA:H_2O	LC-Orbitrap	Gregory et al., 2012
Tissue	CE, GP, GL, SL, FA	1:1 1-butanol:MeOH	Direct injection	LC-ESI-MS/MS	Alshehry et al., 2015
Tissue	AC	8:6:7 MTBE:MeOH:H_2O (MTBE fraction)	65:30:5 ACN:IPA:H_2O	LC-Orbitrap	S. Chen et al., 2013; Y. Chen et al., 2013
Seeds, tissue	FA	2:1 DCM:MeOH (DCM fraction)	Hexane	GC-FID	Cequier-Sánchez et al., 2008
Plant	CE	2:2:1.2 CHF:MeOH:H_2O (CHF fraction)	CDF	NMR	Brasili et al., 2014
Oil	SH, H, ST	Hexane; fractionation	Isooctane or toluene	GC–MS	Alizadeh et al., 2017
Oil	SH, MA, TP, DI	Hexane; fractionation	DCM	GCxGC-TOF	Casilli et al., 2014
Oil	SH, TP	9:1 DCM:MeOH	Direct injection	GCxGC-FID	Gros et al., 2014
Oil	SH, TP	9:1 DCM:MeOH; fractionation	DCM/methanol	GCxGC-TOF	Hall et al., 2013
Oil	Heteroatom	toluene	Direct injection	FT-ICR/MS	Hur et al., 2018
Polar extractions					
Wastewater, sewage, sediment	OC	1:1 MeOH:ACN	Direct injection	LC-TOF	Rostkowski et al., 2013
Sediment	Unknown	Cell lysis; cold MeOH	Pyridine	GC–MS	Beale et al., 2017
Dissolved organic matter	LN, OS, TN, SO	MeOH	Direct injection	FT-ICR/MS	D'Andrilli et al., 2013
Dust, Lint	PH, AA, OC, FA	MeOH then ACN	1:1 MeOH:ACN	LCxLC-TOF	Ouyang et al., 2017
Soil	OS, AA, NT	Fumigation, direct CHF-K_2SO_4, cold MeOH:CHF:H_2O, & hot EtOH	100 mM NH_4HCO_2 in 25% ACN	GC–MS	Warren, 2015
Soil	OS, AA, NT	H_2O, 10 mM K_2SO_4, 10 mM NH_4HCO_3, IPA:MeOH:H_2O, & 10–100% MeOH	Pyridine	GC–MS	Swenson et al., 2015
Soil biocrust	OS, AA, NT	H_2O	MeOH	LC-Orbitrap & LC-QTOF	Swenson et al., 2018
Soil	LN, OS, TN, SO	H_2O	MeOH	FT-ICR/MS	Jiménez-Morillo et al., 2018
Tissue	FA	8:06:07 MTBE:MeOH:H_2O (MeOH/H_2O fraction)	1:4 MeOH:H_2O	LC-QTOF	S. Chen et al., 2013; Y. Chen et al., 2013
Plant	AA, OS	2:2:1.2 CHF:MeOH:H_2O (MeOH/H_2O fraction)	1:2 CD_3OD:D_2O	NMR	Brasili et al., 2014

[a] Classes include: fatty acids (FA), oligosaccharides (OS), glycerophospholipids (GP), glycerolipids (GL), sphingolipids (SL), bile acids (BA), cholesterol-derivatives (CE), phenol lipids (PL), amino acids (AA), proteins (P), acylcarnitines (AC), organic contaminants (OC), saturated hydrocarbons (SH), monoaromatic steroids (MA), thiophenes (TP), diamondoids (DI), alcohols (OH), flavonoid (Fl), cyclitol (CY), phenolic acid (PA), hydroxycinnamic acids (HA), polyaromatic hydrocarbons (PAH), squalene (SQ), biohopanepolyols (BHP), hopanes (H), steranes (ST), chlorins (CHL), nucleotides (NT), tannins (TN), lignin (LN), soot (SO), phthalate (PH).
[b] Solvents include: dichloromethane (DCM), methyl-tert-butyl ether (MTBE), methanol (MeOH), ethanol (EtOH), acetonitrile (ACN), hexafluoroisopropanol (HFIP), chloroform (CHF), isopropylalcohol (IPA), water (H_2O), deuterated methanol (CD_3OD), deuterated water (D_2O), deuterated chloroform (CDF).
[c] Analytical instrumentation includes: gas chromatography(GC), two-dimensional gas chromatography (GCxGC), liquid chromatography (LC), two-dimensional liquid chromatography (LCxLC), time of flight mass spectrometer (TOF), quadrupole time of flight mass spectrometer (QTOF), nuclear magnetic resonance spectroscopy (NMR), mass spectrometer (MS), tandem mass spectrometer (MS^n), flame ionization detector (FID), Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR/MS).

contamination incorporated during the extraction and need to be incorporated into the initial experimental design. As a rule of thumb, the number of method blanks should be 10% of the total number of samples including replicates.

- **Recovery standards** can be used to measure the extraction consistency by adding known amounts of an artificial (i.e. isotope labelled) compound into the sample before extraction to determine how much is lost during extraction.

Aside from QC procedures, sample and extract storage are important to consider. Organic molecules will degrade after sampling due to light, oxygen, and temperature deviations from the samples' native environment. The use of certain solvents, like methanol, can generate artifacts during sample extraction and storage (Sauerschnig et al., 2017). Thus, long-term storage of samples and extracts should not be at room temperature unless freeze-dried or evaporated, and as much as possible, analyses should proceed immediately after extraction.

It is also important to sample and extract with other interferents in mind. Only use high-grade solvents for extraction and analysis to avoid the introduction of contaminants. Avoid plastic as much as possible. Solvents like dichloromethane and chloroform can strip plasticizers from plastic during extraction, and non-polar compounds have a strong adsorption capacity for plastics. Contamination from plastic manufacturing can be highly variable, even within plastic consumables from the same lot (Yao et al., 2016). At a minimum, dichloromethane and/or chloroform resistant plastics should be used, but the use of glass will increase yields and reduce the interference from plasticizers. If plastic use is unavoidable, which it so often is, proper method blanks can account for some of the plastic-derivatives, but not all (Yao et al., 2016). The highly sensitive instrumentation used in untargeted "-omics" studies can detect trace levels of plastic, which can impact later data analysis.

3. Analysis

3.1. Detection

3.1.1. High Resolution Mass Spectrometry

High resolution mass spectrometry (HRMS) detection is the work horse of untargeted "-omics" methods because of its resolving power; but how resolution is defined is important, and it can be defined in different ways (Hernández et al., 2012). Here resolution will be defined using full width half maximum, which is the manufacturing standard

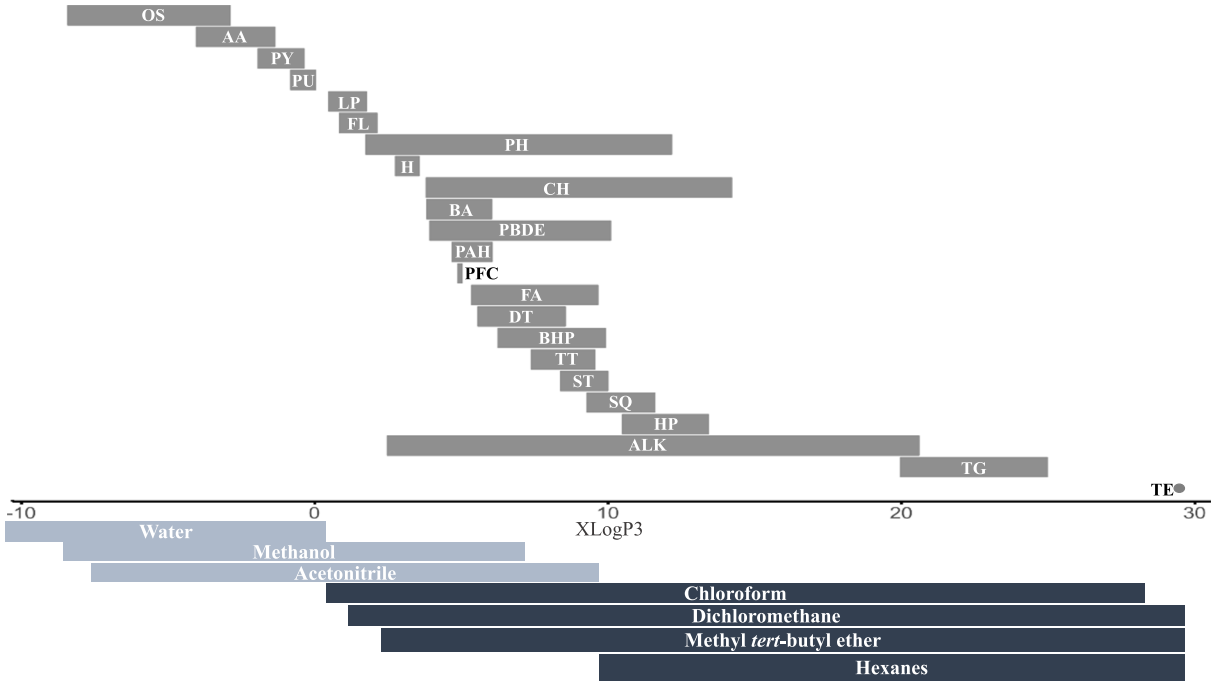


Fig. 3. Solvent and common biomarkers chart. Each family of compounds is represented by a range of XLogP3 (calculated octanol/water coefficients where <0 is more polar and >0 is more non-polar) values found in the PubMed Database (Kim et al., 2016). The families do not contain all molecules within the family but rather a subsample chosen by relevance and different polarities to maximize range. Below the plot are the common solvents for "-omics" studies, and their approximate range of extractable metabolites. If the solvents overlap, then they are miscible. [a]Compound shortforms are: oligosaccharides (OS), amino acids (AA), pyrimidines (PY), purines (PU), lignin phenols (LP), flavonoids (FL), phthalates (PH), hormones (H), chlorins (CH), bile acids (BA), polybrominated diphenyl ethers (PBDE), polyaromatic hydrocarbons (PAH), polyfluorinated compounds (PFC), fatty acids (FA), diterpenoids (DT), bacteriohopanepolyols (BHP), triterpenoids (TT), sterols (ST), squalene-derivatives (SQ), saturated hopanes (HP), saturated alkanes (ALK), triacylglycerides (TG), and glycerol dialkyl glycerol tetraethers (TE).

(Murray et al., 2013). The minimum resolution for an HRMS instrument is 10,000. Instruments with this ability include: (1) magnetic sector (10,000–80,000 resolution), (2) time of flight (TOF; 20,000–80,000 resolution), (3) orbitrap (up to 500,000 resolution at 200 *m/z*), and (4) Fourier-transform ion cyclotron resonance (FT-ICR; up to 1,000,000 resolution; Hernández et al., 2012; Krauss et al., 2010; Špánik and Machyňáková, 2018). Most HRMS instruments also have high mass accuracy (±0.001 Da), wide mass ranges (up to 2000 Da), and high sensitivity (Hollender et al., 2017; Krauss et al., 2010). What this translates to for untargeted "-omics" analysis is: (1) high resolution allows the MS to separate more compounds from complex matrices, (2) high mass accuracy allows structural information to be determined, and (3) large mass ranges allows analysis of larger molecules/polymers. While low resolution mass spectrometers (LRMS) have been used in the past, next-generation HRMS instruments are preferred and add unparalleled information to "-omics" studies.

An important consideration of HRMS is the choice of ionization technique. There are several types of ionization such as: (1) electrospray ionization (ESI), (2) atmospheric pressure photoionization (APPI), (3) electron impact (EI), and (4) atmospheric pressure chemical ionization (APCI). Ionization methods are grouped into hard (EI), and soft (ESI, APCI, APPI). Typically, HRMS "-omics" studies use soft methods because it provides key information for structural identification if the mass of the entire compound is known. However, no single ionization technique is good for all compounds in all situations. Using multiple ionization methods obtains the most comprehensive information, but this is not always possible due to cost restrictions and instrument compatibilities. The advantages and disadvantages of the different ionization methods have been previously discussed (Cai and Syage, 2006; Gutiérrez Sama et al., 2018; Imbert et al., 2012).

A new feature for HRMS instruments is the addition of ion mobility mass spectrometry (IMMS). Typical HRMS instruments only sort by mass and charge, IMMS separates ions by their size, shape, charge, and mass. Recently, there have been a number of reviews regarding IMMS and "-omics" applications (Cajka and Fiehn, 2016; Hinz et al., 2018; Mairinger et al., 2018; Zhang et al., 2018). The application of IMMS could: (1) reduce background noise, (2) increase selectivity, (3) discriminate isomeric molecules - molecules with the same chemical formula but different structures, and (4) add structural identification information (Cajka and Fiehn, 2016; Lanucara et al., 2014). Collision cross-section (CCS) values are of particular importance. CCS values represent the time the ion requires to cross the chamber, which is dependent on ion shape (Lanucara et al., 2014). CCS values are unique to compounds, and have high inter-laboratory reproducibility as setting changes in mass spectrometer or hyphenated chromatography instruments do not influence the value (Lanucara et al., 2014; Paglia et al., 2014, 2015). This kind of technology has been applied to complex mixtures such as petroleum samples (Fernandez-Lima et al., 2009; Gutiérrez Sama et al., 2018), lipidomics (Hinz et al., 2018; Paglia et al., 2015; Paglia and Astarita, 2017; Poad et al., 2018), and metabolomics (Dwivedi et al., 2010; Paglia et al., 2014; Paglia and Astarita, 2017).

3.1.2. Nuclear Magnetic Resonance (NMR)

NMR is another routine detection technique for small molecule "-omics" methods. NMR has several advantages over mass spectrometry: (1) it is non-destructive; (2) reproducible; (3) provides more structural information; (4) sample preparation is simple; (5) does not produce matrix effects; (6) discriminates isobaric compounds; and (7) no derivatization is required (Gathungu et al., 2018; Li et al., 2017). Yet despite its many advantages, NMR is less frequently used than mass spectrometry for several reasons: (1) less sensitivity; (2) overlapping signals; (3) larger samples needed especially for carbon NMR (^{13}C NMR); (4) fewer databases for structural annotation; and (5) poor resolution of compounds in the same class (Li et al., 2017).

However, NMR use is expanding, and new technology is being developed to mitigate its disadvantages. Spectroscopic analyses like two-dimensional proton-proton homonuclear total correlation spectroscopy (2D-^{1}H, ^{1}H TOCSY) or two-dimensional proton-carbon heteronuclear

single quantum coherence spectroscopy (2D-^{1}H,^{13}C HSQC) can increase compound resolution (Heude et al., 2017). In addition, LC hyphenated NMR techniques are in development, which will hopefully lead to better compound resolution, and spectral decongestion (Gathungu et al., 2018). NMR has been used on soil samples to assess the effects of environmental gradients (Pisani et al., 2016), soil physical states (Masoom et al., 2016), plant tissue (Brasili et al., 2014) and soil organic matter turnover (Wang et al., 2017).

3.2. Hyphenated techniques

3.2.1. Liquid Chromatography Mass Spectrometry (LCMS)

LCMS is the most commonly used hyphenated chromatography technique. LC separates compounds before they enter the mass spectrometer. With LC there is better sensitivity, compound resolution, and reduced ion suppression than direct injection onto the MS alone (Cajka and Fiehn, 2016).

How to separate compounds with LC is an important consideration for method development. There are several separation techniques for LCMS: (1) reverse phase LCMS, (2) normal phase LCMS, and (3) hydrophobic interaction liquid chromatography (HILIC) - a variant of normal phase LCMS (Cajka and Fiehn, 2016). The use of reverse phase LCMS provides separation based on lipophilicity with polar compounds separating best (Cajka and Fiehn, 2014, 2016). HILIC and normal phase LCMS both separate compounds based on hydrophilicity. Normal phase LC has been applied to petroleum samples and is extensively reviewed by Kamínski et al. (2005). HILIC is better known to separate small polar molecules (Buszewski and Noga, 2012; Nováková and Vlčková, 2009). Thus, the choice of extraction method (i.e. polar or non-polar) will influence the type of LC separation technique used.

Recently, the use of two-dimensional liquid chromatography (LCxLC) with HRMS has been increasing. There are two reasons to use LCxLC: (1) Matrices are too complex for one-dimensional liquid chromatography, and (2) isomeric/isobaric species of interest require increased resolution (Cajka and Fiehn, 2016; Stoll and Carr, 2017). LCxLC methods have been demonstrated for untargeted analyses of blood plasma (Holčapek et al., 2015; Wang et al., 2013), dust and dryer lint (Ouyang et al., 2017), wastewater (Haun et al., 2013; Ouyang et al., 2015), and rice (Navarro-Reig et al., 2018).

3.2.2. Gas Chromatography Mass Spectrometry (GCMS)

GCMS is another useful hyphenated technique for "-omics" methods. GC paired with HRMS is a powerful tool in untargeted analyses according to a recent review by Špánik and Machyňáková (2018). GCMS is ideal for volatile and semi-volatile compounds, with the option of modifying non-volatile compounds to increase GCMS compatibility. Thus, GCMS can analyze a wide range of compounds for "-omics" studies.

A caveat of GCMS is the ionization techniques. Commonly, GCMS is paired with EI ionization. Extensive compound libraries exist for EI fragmentation patterns; however, since EI is a hard ionization source, it is rare that the compound mass can be determined. This is a disadvantage when working with unknown compounds. Thus, for untargeted analysis, having a GCMS that is also capable of using soft ionization (i.e. chemical ionization, APCI, APPI, etc.) is useful for structural annotation, but soft ion ionization methods can also be less sensitive (Tranchida et al., 2018). The application of GCMS is extensive and includes the untargeted analysis of petrochemicals (Gros et al., 2014; Rodgers and McKenna, 2011), migrating compounds from packaging (Cherta et al., 2015), and environmental contaminants (Hernández et al., 2012; Serrano et al., 2011).

Recently, the popularity of comprehensive two-dimensional gas chromatography (GCxGC) with HRMS has grown immensely. A number of reviews have been written about GCxGC (Mondello et al., 2008), and to describe its applications to petrochemicals (Pollo et al., 2018), environmental analyses (Pani and Górecki, 2006), and forensic analyses (Gruber et al., 2018). GCxGC with LRMS can provide better sensitivity than GC HRMS alone (Fushimi et al., 2012), but it becomes more powerful when the separation efficiency of GCxGC is combined with HRMS. Parastar et al. (2018) determined that GCxGC-TOF far surpassed GCMS in resolution and sensitivity. GCxGC-HRMS has been used to screen samples for halogenated compounds (Ieda et al., 2011; Pena-Abaurrea et al., 2014), petrochemicals (Casilli et al., 2014; Hall et al., 2013; Rodgers and McKenna, 2011; Walters et al., 2018), dust constituents (Hilton et al., 2010), and sewage sludge (Veenaas and Haglund, 2017). Thus, due to the outstanding capabilities of GCxGC technology there has been increasing interest in using it for untargeted "-omics" work.

3.3. Analysis quality control

Even with HRMS instruments, analysis quality control (QC) is integral to reducing analytical variability. The purpose of analysis QC is to reduce the effects of: (1) Batch Effect - shifts in retention time and intensity between batches of samples analyzed at different times or days leading to data processing issues (Rusilowicz et al., 2016; Wehrens et al., 2016), (2) Matrix Effect – where co-eluting compounds, buffer additives, inadequate flow rate, poor solvent purity, and an unclean source can lead to ion suppression/enhancement, reduced sensitivity, quantification/qualification problems, and non-detection (Furey et al., 2013; Uclés et al., 2017), (3) Carryover – when analytes from the previous sample appear in the following sample leading to poor quantification/qualification, and loss of accuracy (Dudzik et al., 2018). It is necessary to consider these effects as they can seriously impact data analysis and may prevent the integration of data from multiple batches and/or experiments. Below are the potential analysis QC's, and which effect they mitigate:

- **Pooled QCs** are samples containing aliquots from multiple samples pooled together and then extracted following the same procedure, or post-extraction aliquots pooled together. Pooled QCs are used to equilibrate the instrument to the sample matrix before running samples and can correct for batch effect (Dudzik et al., 2018; Dunn et al., 2012).
- **Randomization** of the analysis order is critical to also reduce bias introduced by batch effect and carryover (Dudzik et al., 2018; Dunn et al., 2012).
- **Blanks** are mobile phase blanks and sample solvent blanks. Running the same extraction/analysis solvent allows the removal of compounds originating from the solvents later during data analysis, and helps monitor for carryover when a set of intermittently placed blanks are throughout the run (i.e. approximately every 10 samples depending on the sample matrix).
- **Cleaning Blanks** are blanks containing strong solvents (i.e. like isopropyl alcohol) to clean the system during analysis and reduce carryover effect.
- **Internal Standard** is a compound (often isotopically-labelled) that is spiked into post-extraction samples and pooled QCs. The internal standard, and recovery standard if used, are also analyzed without the matrix, but in the same solvent as the samples. The internal standard controls for variations in solvation and injection volumes between samples. The magnitude of the matrix effect is determined by comparing the intensity of the spiked samples to just the internal standard in the sample solvent. Matrix effect is a percentage (intensity of spiked internal standard in sample/internal standard in blank solvent x 100%). The strength of the matrix effect can be grouped into three categories: (1) No Matrix Effect (|0|–|20|%), 2) Weak Matrix Effect (|20|–|50|%), and 3) Strong Matrix Effect (>|50|%; Uclés et al., 2017).
- **Technical Replicates** are repeated analyses of the same samples and are not to be confused with environmental or sampling replicates. Each sample should have six replicates which consists of any combination of environmental, sampling, and technical replicates. The replicates should be randomized so that they are all not done in the same

batch or one after another. A large number of replicates mitigates batch effect and carryover influences, which will make data integration easier, and also lend more statistical power to data analyses (Bader et al., 2016).

4. Data mining

4.1. Tools for data mining

Data mining is the process of handling and analyzing data in such a manner as to "mine" the information buried within. Currently, there is a large push to move data mining in metabolomics and lipidomics into simple, but flexible, data workflows that increase reproducibility, and are easily accessible. We also advocate such data workflows for sediment-"omics" and petroleomics. Cambiaghi et al. (2016), Gorrochategui et al. (2016), Weber et al. (2017), and Yi et al. (2016) have all reviewed data workflows for "-omics". It should be noted that, we are focussing on untargeted "-omics" data workflows.

There are many tools to data mine "-omics" data. Spicer et al. (2017) wrote a comprehensive review of open-source tools that can be found to process metabolomics data. Websites hosted by OMICtools (https://omictools.com) and ms-utils (http://www.ms-utils.org) maintain updated lists of available tools along with forums and relevant references. Briefly, common commercial tools include: (1) Progenesis QI (Waters), (2) Compound Discoverer (Thermofisher), (3) XCMS^{plus} (ABSciex), (4) MetaboScape (Bruker), and (5) MassHunter (Agilent). There are also a number of open-source workflows: (1) Workflow4metabolomics (Giacomoni et al., 2015), (2) Galaxy-M (Davidson et al., 2016), (3) XCMS (Tautenhahn et al., 2012b), (4) MetaboAnalyst 3 (Chong et al., 2018), (5) MAVEN (Melamud et al., 2010), (6) MAIT (Fernández-Albert et al., 2014), (7) SECIMTools (Kirpich et al., 2018), and (8) MZmine2 (Pluskal et al., 2010). A research survey hosted by Weber et al. (2017) determined that most LCMS users use open-source software like XCMS and MZmine2. GCMS users were split between commercial software and open-source. Other useful resources are how-to publications detailing MZmine2 (Verkh et al., 2018), XCMS (Forsberg et al., 2018), and MetaboAnalyst (Xia and Wishart, 2011, 2016) workflows.

There are many different tools, and it is easy to become paralyzed by the number of options available. In addition, there are major difficulties with tool compatibility where the output of one tool is not acceptable to the next tool (Spicer et al., 2017). Thus, there has been a major push to generate open-source tools that contain the entire workflow, but are also flexible to address different datasets and transparent to facilitate reproducibility. Thus, learning with an open-source workflow is often a good place to begin.

4.2. Data preprocessing

4.2.1. File types

The first step in data mining is knowing where your data are coming from, and how your data were acquired. The user can acquire HRMS data in centroid mode or profile mode. Profile mode looks like a Gaussian curve across *m/z* values but varying in intensity. Centroid mode is a compressed profile mode where only the *m/z* values associated with the highest intensity are reported. The choice of acquisition mode influences the resulting data. Centroided data files are smaller than profiled data, but result in the loss of noise characteristics, ion signal linearity, interfering ions, and informative isotope features (Wang and Gu, 2010).

Additionally, most of the vendors have proprietary file types. The difficulty with propriety file types arises when a researcher does not have access to the vendor's commercial software or would like to perform a statistical analysis that is beyond the scope of their software. The researcher then has a difficult to open file. If a researcher is looking to use a specific instrument for "-omics" analysis, then it is prudent to also ask the vendor in question how to convert their files to more open-access file types like mzXML, mxData (Orchard et al., 2007), mzML (Martens et al., 2011), netCDF (Gorrochategui et al., 2016). There are also available open-source software packages, such as Proteowizard (Holman et al., 2015) to convert proprietary file types into open-access files.

4.2.2. Spectral processing

The processing of raw HRMS files takes several steps and each software package will have its own algorithms, or have multiple algorithms to choose from (Gorrochategui et al., 2016). Thus, reviewing the advantages and disadvantages of each algorithm is useful, but often the data dictate which algorithms to use. Reviews of this process are available in Gorrochategui et al. (2016), and Karaman (2017).

Software choice will affect the results (Coble and Fraga, 2014; Gürdeniz et al., 2012). Studies have compared the resulting biomarkers from different software programs and discovered only two of fourteen potential biomarkers were common between them (S. Chen et al., 2013; Y. Chen et al., 2013). A recent paper compared MZmine2 and XCMS and suggested problems within their algorithms (Myers et al., 2017a); however, in a following paper they implemented an improved algorithm that also minimizes false discovery rates (Myers et al., 2017b). This is not to say that open-source software is of lesser quality than commercial software. X.M. Li et al. (2018) and Z. Li et al. (2018) determined MZmine2 outperformed other open-source software and commercial software with the lowest false discovery rate and highest identification of true positives. It is an advantage for open-access software that the research community can assess their accuracy and precision, which ultimately benefits the field as a whole. Thus, the user should be aware that their choice of commercial or open-source programs will impact their results, but this is a rapidly developing field and open-source offers transparency.

4.2.3. Missing value imputation and data normalization

After spectra processing the dataset then needs to be curated to move forward into the statistical analysis. The following steps are flexible and can be used in any combination thereof; however, there is a purpose to each step and the result should be a dataset following a typical Gaussian distribution.

The first technique is missing value imputation. Missing values in a dataset can cause problems during the statistical analysis (Di Guida et al., 2016), and are usually attributed to non-detects, peak picking errors during spectral processing or values below the analytical noise threshold. Some missing values can be addressed by gap filling algorithms to fill in small or missing peaks (Karaman, 2017). Missing value imputation algorithms are applied if missing values persist. Wei et al. (2018) proposed several methods to address missing values. Simple filtering procedures can be applied where only bins (i.e. retention time + *m/z* variables) with >80% detection frequency in any one sample group are kept (Yang et al., 2015) or removing samples/replicates with a high proportion of missing values (Armitage et al., 2015). Imputation algorithms can be applied as well, either in lieu of or after filtering, and studies have compared the impact of the different imputation strategies (Armitage et al., 2015; Di Guida et al., 2016; Gromski et al., 2014b; Wei et al., 2018). The purpose of this strategy is to reduce the number of "zeros" in the dataset, which improves the performance of many different statistical analyses.

After the missing value imputation, row-wise corrections (i.e. within chromatogram corrections), also called normalization, are performed. The purpose of normalization is to correct for systemic bias in ion intensities from sample collection, experimental bias, and analytical variations (Gorrochategui et al., 2016). There are several reviews discussing different normalization strategies (De Livera et al., 2012; Li et al., 2016; Chen et al., 2017). Each normalization strategy has its own advantages and disadvantages (Karaman, 2017). The internal, external, or pooled QC's, and the data structure itself determines the normalization strategy used. Every dataset is different and will require different

techniques as demonstrated by Wu and Li (2016) who compared the results of different normalization procedures on two biological matrices. Usually, multiple normalization strategies are compared. The success of the normalization is judged by a Gaussian distribution of intensities, and reduced distance within sample groups combined with increased distances between groups on a principal component analysis (PCA) - described later.

Lastly, data processing requires column-wise corrections (i.e. between chromatograms), which involves data scaling and transformations. The purpose of scaling is to correct for fold changes that could mask important lower intensity peaks (Van den Berg et al., 2006). The most commonly used scaling methods are autoscaling (Kvalhein, 1985), and pareto-scaling (Kasprzak and Lewis, 2001). The purpose of transformations is to correct for heteroscedasticity (also known as unequal variance due to noise) and skewness. Many statistical analyses require that analytical noise is homoscedastic (Karaman, 2017). Van den Berg et al. (2006), Di Guida et al. (2016) and Karaman (2017) compare the advantages and disadvantages of scaling and transformation methods.

4.3. Statistical analysis

The 2D matrices generated from the spectral processing can be analyzed with many different statistical analyses. Typical "omics" workflows use discriminatory analyses to generate hypotheses about which molecular compounds are potential biomarkers. These discriminatory analyses include univariate and multivariate statistical tests such as ANOVAs with multiple pairwise comparisons, volcano plots, partial least squares (PLS), random forest (RF), and support vector machines (SVM) to list a few. In addition, there are many pattern recognition algorithms that provide useful information such as molecular fingerprinting, unsupervised multivariate analyses, and clustering. A single data set does not require all of these analyses, and the choice of test is data dependent. Often pattern recognition algorithms assess data structure and group differences. Discriminatory analyses determine what molecules are driving variance between groups. Here we present commonly used statistical tests for "omics" data analysis.

4.3.1. Molecular fingerprinting analysis

Molecular fingerprinting analyses visualize differences in samples. These analyses require the use of HRMS, typically FT-ICR or Orbitrap. The elemental compositions (i.e. molecular formula) are estimated based on the accurate mass measurements, isotopic ratios and atomic limitations based on the type of molecules expected in the sample (Section 5.2; Kind and Fiehn, 2007; Stubbins et al., 2010). Once the molecular formulas have been estimated, then double bond equivalents (DBE) calculations, Kendrick mass defect plots, van Krevelen diagrams or multidimensional stoichiometric compound classifications can be built (Gutiérrez Sama et al., 2018; Rivas-Ubach et al., 2018; van Krevelen, 1950). Molecular fingerprinting is commonly used in petroinformatics (Gutiérrez Sama et al., 2018; Hur et al., 2017, 2018), but has been applied to soils (D'Andrilli et al., 2013; Jiménez-Morillo et al., 2018).

4.3.2. Univariate analysis

Several univariate analyses can be performed on "-omics" data. Univariate analyses consider only one dependent variable, whereas multivariate considers more than one. Common univariate tests include: Fold change, *t*-tests, volcano plots, correlation analysis, analysis of variance (ANOVA), Kruskal-Wallis and/or multiple comparison.

Univariate analyses require establishing whether parametric (e.g. Pearson correlation, ANOVA, *t*-tests, Tukey HSD, or Fisher's LSD) or non-parametric (e.g. Mann–Whitney *U* test, Spearman correlation, Kruskal-Wallis, Dunn's test) tests are appropriate. According to Pinto (2017), non-parametric tests are less powerful, but only when there are normal populations, equal variances (i.e. homoscedastic) and equal sample sizes. Thus, the choice of parametric or non-parametric strategies is entirely dependent on the data.

Fold changes indicate absolute value changes between two group means (i.e. whether one group demonstrates an increase or decrease in magnitude when compared to a base group or control). *t*-tests demonstrate whether a variable is significantly different between two groups. Volcano plots compare two groups using statistical significance (y-axis) and fold change (x-axis). A "volcano" shaped arrangement of points often appears whereby points further from 0 on the x-axis demonstrate greater fold change, and points higher on the y-axis are more significant (i.e. on a -log(*p*-value) scale y-values > 1.3 indicate *p*-values < 0.05; Hur et al., 2017). The main limitation of fold changes, *t*-tests, and volcano plots is that they only compare two groups.

Multi-group analyses compare more than two groups and include ANOVA, Kruskal-Wallis, correlation analysis and multiple comparisons. ANOVA (parametric) and Kruskal-Wallis (non-parametric) compare means between more than two groups. They indicate if there are differences between the groups and assign *p*-values. Multiple comparison tests (e.g. Tukey HSD, Fisher's LSD, Dunn's test, etc.) inform about where the differences lie between the groups. (Pinto, 2017; Xi et al., 2014). Lastly, correlation analysis can also be used for both two group, and multi-group analyses. Correlation analyses visualize the degree of association between variables, and can even be combined with molecular fingerprinting analyses performed above (Asuero et al., 2006; Hur et al., 2017). Multi-group analyses are well established methods, however for large data sets or when there are many groups, they can be computationally intensive.

As a side note, it is common in "-omics" for the number of measured variables to be far greater than the number of samples due to the nature of the data (e.g. there are thousands of organic molecules in one sample). This imbalance increases the risk of false discoveries and post-analytical variation because when there are too few samples the variation from the variables is over-emphasized (Martins et al., 2018). Thus, when using multiple comparisons, (i.e. post-hoc tests) the *p*-values often have to be corrected using Bonferroni, Benjamini–Hochberg, or metabolome-wide significance level corrections (Pinto, 2017; Xi et al., 2014). If corrections cannot be applied then the results should be interpreted with caution.

4.3.3. Multivariate analysis

Multivariate analyses are used to determine relationships between multiple dependent variables, and are better at handling batch-to-batch variation and drift (Pinto, 2017). There are two main categories: (1) Unsupervised - where no assumptions are drawn from group classifications (i.e. Principal component analysis (PCA), hierarchical clustering (HCA) and self-organizing map (SOM)), or (2) Supervised - where groups and samples are associated together to build the final model (e.g. partial least squares or projection to latent structures (PLS), random forest (RF), and support vector machines (SVM)). Unsupervised models identify outliers, and determine which variables are the most influential without group bias. Often an unsupervised model can determine whether variance from pre-analytical, analytical, and post-analytical artifacts are overwhelming the desired biological/experimental variance. Supervised methods are powerful models that integrate groupings (i.e. impacted or unimpacted) into the generated matrices so that the model correctly associates the research questions with the output (Goodacre et al., 2004). Supervised models isolate candidate biomarkers from the data in discriminant analyses, as demonstrated in a S-plot derived from an orthogonal PLS (Fig. 4C). Usually, variables of importance (VIP) or selectivity ratios can be displayed to illustrate which compound signals are the most discriminatory (Farrés et al., 2015b). Below, the analyses are briefly explored.

PCA models explain variance within the data set. A PCA generates a scores plot (Fig. 4A) – comparing separation between samples, a loadings plot – expressing direction of variance for each variable (i.e. detected compounds), and a scree plot – showing the percent variance

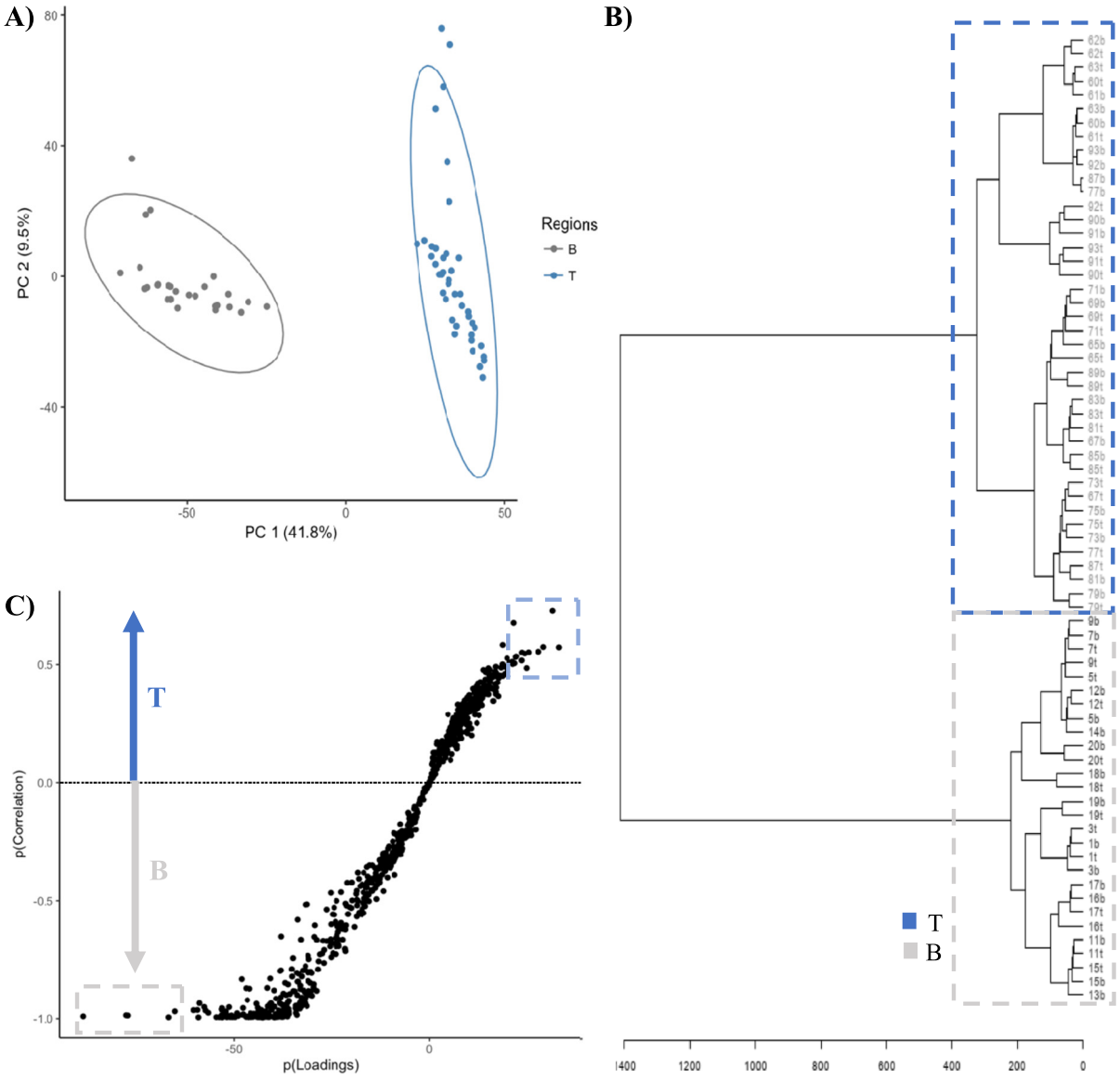


Fig. 4. Example outputs of the chemical similarity between lakes based on their molecular composition. A) Principal Component Analysis two-dimensional scores plot - similar samples are clustered closer together. Principal component 1 (PC1) explains 41.8% of the total variance and defines the component where groups of the samples are most dissimilar. PC2 explains 9.5% of the variance. Overall this model explains 51.3% of the variance in the first two dimensions. B) Hierarchical Clustering dendrogram based on Euclidean distance and the Ward linkage algorithm. Ideally, the samples classed together should appear closer together based on the linkages in the dendrogram, which indicates similarity (highlighted in boxes). C) S-plot is a loadings scatter plot for a discriminant OPLSDA model. The y-axis represents the correlation where the increasing |+/-| distance from 0 indicates the more likely the variable (i.e. unique mass + retention time signal) occurs in a class (i.e. either T or B in this case). The x-axis represents the covariance and indicates the variable's magnitude of contribution to the model. Thus, the variables that are the most reliable and contribute the most to the model are at the extremes of the ideal "S" shape (highlighted in boxes). These variables are discriminatory for each class (i.e. T or B).

explained by each principal component. PCA allows for visual interpretation of the variance. Samples grouped close together are more similar than those grouped farther apart. The loadings plot is useful in determining which variables are key to defining the variance in any one principal component. Using a PCA, outliers can be identified as they often do not cluster with the others, they are outside a 95% confidence window, using cumulative measures of distance, or by inspecting model residuals (Pinto, 2017). PCA is the work horse of multivariate data analyses and provides insight into the data structure.

HCA is an unsupervised method, and it is used to visualize natural clustering in the data with dendrograms as demonstrated in Fig. 4B (Ren et al., 2015). Distance measures (e.g. Euclidean distance, Pearson's correlation, Spearman's rank correlation) and linkage functions (e.g. average linkage, complete linkage, or Ward's linkage) are chosen to build a dendrogram. Useful reviews for deciding on the best combination can be found in Hastie et al. (2009), Jain et al. (1999), and Ren et al. (2015). In addition, there are several papers comparing distance measures, which include methods for assessing the model validity (Giancarlo et al., 2010, 2011; Jaskowiak et al., 2013, 2014). The major advantage of HCA is that it is easy to visualize, but interpretation can be difficult when there are a large number of samples or too many variables (Pinto, 2017).

Another unsupervised analysis is SOM. SOM is a neural network algorithm that visualizes multi-dimensional data in several ways including using heatmaps, Umatrix, and HCA. SOM is a type of artificial neural network that uses neighbourhood functions to arrange "neurons" or nodes in a grid (Ren et al., 2015). Thus far there are only a few examples available in literature involving the use of SOM (Binder and Wirth, 2015; Franceschi and Wehrens, 2014; Haddad et al., 2009).

PLS is a common statistical analysis for "-omics" work. It is a supervised regression method that can be used to maximize the separation between sample groups (Xi et al., 2014). There are different versions

of PLS: orthogonal PLS (OPLS; Thévenot et al., 2015), and sparse PLS (sPLS; Lê Cao et al., 2011). Presently, PLS methods are poorly utilized as many researchers do not optimize or correctly report parameters (Brereton and Lloyd, 2014; Considine et al., 2018). PLS models are also prone to over-fitting where the results cannot be replicated with a new set of data (Westerhuis et al., 2008; Xi et al., 2014). Thus, the models need to be validated to avoid overfitting and proper parameter optimization. There are several ways to validate PLS models including cross-validation, Monte Carlo cross-validation (MCCV), receiver operating characteristic (ROC) analysis, and permutation tests (Pinto, 2017; Westerhuis et al., 2008; Xi et al., 2014). In addition, a high Q2 value (commonly enlisted to indicate model quality) is not always indicative of a good model, albeit it is generally a good indicator of a poor one, so care should be taken during the interpretation as well (Gromski et al., 2015; Westerhuis et al., 2008).

Random forest is a supervised learning algorithm that is useful for classification, identification of outliers and discriminant analysis of biomarkers (Breiman, 2001; Cutler et al., 2007). It is adept at handling large datasets, missing values, resistant to outliers and over-fitting, and it can handle noisy data (Gromski et al., 2015). A useful metric for RF analyses are out-of-bag (OOB) errors, where a lower error indicates a better model. OOB errors are calculated during the RF computation so no additional validation methods are required (Kuznetsova, 2014). A current issue with RF is the lack of simplistic visualization (Gromski et al., 2015), although visualization techniques are under development (da Silva et al., 2017; Kuznetsova, 2014).

The last method discussed here, is the supervised SVM method. SVM is a nonparametric machine learning method that can be used for classification and regression (Gromski et al., 2015). This method is not influenced by population distributions, and is resistant to outliers, overfitting and noise (Gromski et al., 2015; Xu et al., 2006). However, currently, there are also a number of disadvantages such as data visualization, solving only two-group problems (multi-group solutions exist, but require additional algorithms), and that it is computationally expensive (Ren et al., 2015). Despite, these disadvantages, in a two-group dataset, SVM has been shown to outperform PLS in classification accuracy and discriminant analyses (Mahadevan et al., 2008) and is equal to RF (Gromski et al., 2014a).

5. Structure identification

5.1. Reporting standards

Structural identification, or annotation, is a bottleneck in the untargeted "-omics" workflow. A comprehensive review on structural annotation by Dunn et al. (2013) states that in a typical GCMS run, only 50% of the metabolites structures can be identified. In addition, the quality of the identification needs consideration. In 2005, a group of researchers formed the Metabolomics Standards Initiative (MSI: http://www.metabolomics-msi.org), which worked to establish good reporting and data handling methods for metabolomics, but can apply to other "-omics" fields (Fiehn et al., 2007). MSI established the Chemical Analysis Working Group (CAWG). CAWG outlined four levels of identification: (1) Identified compounds – confirmation with an authentic standard with at least two orthogonal approaches; (2) Putative annotated compounds – identification based on physical chemistry and/or spectral libraries; (3) Putatively identified class level – structure remains unknown but physical chemistry and/or spectra suggests a likely class of molecules; and (4) unknown – little to no knowledge on the compound (Sumner et al., 2007). Currently it is expected to annotate compounds of interest according to their level of identification (Creek et al., 2014). The original CAWG reporting standards are sometimes too simple, so other reporting standards have been developed by Schymanski et al. (2014), EU Guideline 2002/657/EC, and Rochat (2017). Identification and/or annotation of compounds can be done during the data acquisition step or after acquisition using additional targeted methods and purification strategies (Dunn et al., 2013). Below, tools and strategies for structural annotation are listed with the caveat that these techniques are typically used in tandem as confirmation from multiple orthogonal techniques provides more confident annotations of the compounds of interest.

5.2. In silico identification tools

There are several different *in silico* tools that facilitate structural annotation. Usually, spectral libraries are integral to the annotation process, but mass fragmenters and elemental compositions calculations are also useful. Below there is a general discussion of tools along with common software.

Common annotation resources are spectral libraries. Spectral libraries provide examples of spectra taken under specific conditions and reference accurate mass, CAS numbers, database specific identifiers, chemical formulas, experimental/predicted physical chemistry data and fragmentation information. There are many databases that are open-access or commercial. Some common open-access databases include:

- **LC-ESI-MS databases** - METLIN (Guijas et al., 2018; Smith et al., 2005; Tautenhahn et al., 2012a), MassBank (Horai et al., 2010), Human Metabolome Database (HMDB; Wishart et al., 2013), LipidMaps (Fahy et al., 2007, 2018), mzCloud, Chemspider (Pence and Williams, 2010), and MetFrag (Ruttkies et al., 2016)
- **NMR databases** - Madison Metabolomics Consortium Database (MMCD; Cui et al., 2008), HMDB, and Complex Mixture Analysis by NMR (COLMAR; Bingol et al., 2012, 2015)
- **GC-EI-MS databases** – Automated Mass Spectral Deconvolution and Identification System (AMDIS), Golm metabolome database (GMD; Kopka et al., 2005).

Useful commercial databases include NIST, and the Wiley Registry. There is also a current movement to develop databases for CCS obtained via IMMS (Zheng et al., 2017). Choose databases with care as the source of the data, the method of acquiring the data, and the focus of the database impacts the quality and/or relevance of the data stored there. Gil de la Fuente et al. (2017) and Vinaixa et al. (2016) provide more extensive reviews of spectral databases.

In silico mass fragmentation is also a handy structural elucidation tool. These tools use algorithms to computationally predict a compound's mass spectrum so that experimental data can be compared with the *in silico* predictions. MetFrag is an example of a reliable open-access fragmenter (Scheubert et al., 2013), and has also been incorporated into the popular METLIN database (Tautenhahn et al., 2012a). For more software and algorithm details please refer to Scheubert et al. (2013).

Elemental composition predictions can be useful for structural annotations and Molecular Fingerprinting Analyses (Section 4.3.1). When acquiring data for elemental composition, there are two important considerations: (1) the mode of acquiring the data, and (2) instrumental accuracy. Profile mode provides better information for determining elemental compositions. In addition, mass accuracy is important. Typically mass accuracy for HRMS instrumentation is 5 ppm (Gorrochategui et al., 2016), but this can vary by instrument and throughout a batch run. However, according to Kind and Fiehn (2006), mass accuracy alone is not enough for accurate elemental compositions and a low (<2%) error for isotopic abundance patterns is also required. There are several open-access algorithms to determine elemental composition such as SIRIUS (Böcker et al., 2009), BRAIN (Claesen et al., 2012), and Fourier (Fernandez-de-Cossio Diaz and Fernandez-de-Cossio, 2012), all of which are recommended by Scheubert et al. (2013). Commercially available software includes: SigmaFit (Bruker), Formula Predictor (Shimadzu), iFit (Waters), and MassHunter (Agilent). Scheubert et al.

(2013) provides more detailed information on the algorithms of the individual tools.

5.3. Analytical identification tools

To have a better annotation for the compounds of interest, often the sample requires further structural elucidation analyses. Many of the *in silico* tools listed above, are useful, but by themselves can only lead to putatively annotated compounds. In addition, there are instances, especially for paleolimnologists and geologists, where standards are unavailable and/or the *in silico* tools do not contain information on the compound. Thus, annotation requires further analytical measurements.

MS can acquire additional information. Accurate masses from HRMS can be reported with the annotation. In addition, tandem mass spectrometry (MS^n; where n = number of mass analyzers) can determine chemical structures. MS^n is a process linking several mass analyzers together to repeatedly filter and fragment compounds. It generates different mass spectral scans that provide additional structural information. MS^n scans are important when trying to discriminate between compounds with the same molecular formula or similar structures (Dunn et al., 2013). Lately, there has been a push to incorporate both HRMS, and simultaneous acquisition of MS/MS (i.e. MS^2) information. For instance, Wrona et al. (2005) suggested an "all-in-one" analysis with a hybrid quadrupole TOF instrument. This kind of all-in-one strategy is called Data Independent Acquisition (DIA), which acquires MS/MS for all *m/z* ions in the specified range (Cajka and Fiehn, 2016). If the reader is interested in DIA instrumentation, Cajka and Fiehn (2016) provide a review on merging untargeted and targeted "-omics" methods.

NMR and infrared spectroscopy (IR) provide additional information when MS and *in silico* methods are not enough. Together MS, NMR, and IR spectra can determine complete structures. MS provides information about the molecular formula using accurate mass and isotopic ratios. NMR demonstrates how are the atoms are connected, and IR provides information about functional groups. To acquire spectra from NMR and IR, however, the compound needs to be purified and concentrated. Packed silica columns, or a preparative LC or GC are used for purification. Also, NMR is typically less sensitive than MS, so greater amounts of the compound are required (Gathungu et al., 2018). Obviously, this process can take a long time, or laboratories may not have access to the instrumentation required. Regardless of data gaps, all the information acquired about an unknown compound should be published. Proper annotation of molecular structures is important to the research community, and high quality spectra of unknown structures can be included in databases, so the information can be made available to other researchers as well.

6. Conclusions

6.1. Challenges

The application of "-omics is an open field for paleolimnologists and geologists alike, but there are challenges facing the wholesale adoption of "-omics". There are challenges in all aspects of the workflow. Presented below are the current challenges for sample preparation, sample analysis, data mining, and structural elucidation.

In sample preparation, there is a lack of consensus and literature on comprehensive OM extraction procedures in environmental matrices. There are many ways to extract OM, but which method is best suited for "-omics"? A common question is, are we extracting too much? What are we missing in this extraction method, and is it important? Sediment-"omics" and petroleomics still require extensive method development. The current literature also does not often include extraction QCs. QCs play a significant role in ensuring the quality of results.

Sample analysis challenges include: expensive instrumentation, and the absence of analysis QCs. HRMS instruments are not cheap, and the fields are reluctant to move away from LRMS instruments like single quad GCMS. Analysis QCs are essential, but also poorly reported.

The entrenchment of older univariate statistical analysis combined with the lack of technical expertise in multivariate analyses stymies the adoption of data mining. Often, a major complaint of sediment"-omics" and petroleomics approaches is that they are too data reliant; however, metabolomics, metagenomics, and transcriptomics fields have demonstrated the reliability of these data workflows. Thus, exposure to these new statistical methods may prove useful to paleolimnologists and geologists alike.

Lastly, one of the largest challenges is annotation. There are no relevant, open-access HRMS and/or NMR libraries for geologically relevant compounds. The absence of libraries severely limits the number of possible structural annotations for biomarker identifications. This represents a major bottleneck in sediment-"omics" and petroleomics because annotation then requires additional analyses (if enough sample is available) and/or extensive, time-consuming literature searches. Standards can also be prohibitively expensive or unavailable so high level annotations may not be possible. In this area there is the potential for massive development in geologically relevant annotation infrastructure.

There is a lot of future work for interested parties, and the fields have a high potential for synergism. Thus, the future of sediment-"omics" and petroleomics requires the development of open-access databases, data repositories and, as a community, discussion on transparency, SOPs, QCs, and practical reporting standards.

6.2. Future implications

"-Omics" methodologies could have large implications for paleolimnological and geological research. "-Omics" is not just a tool for molecular biologists, it is a power screening tool for environmental scientists as well. Untargeted "-omics" presents a comprehensive method for hypothesis generation to add new perspectives to old questions.

There are potentially many applications of "-omics" to paleolimnological and geological samples. Geologists are already beginning to use petroleomics for oil bed discrimination, oil spill reconstructions, and for determining the effects of weathering on crude oil for remediation research. However, petroleomics may benefit from the addition of the data mining tools mentioned here such as additional statistical analyses, QCs, and reporting standards.

In addition, the application of "-omics" to paleolimnology as sediment-"omics", could lead to more biomarkers for paleoecology, archaeology, paleoecotoxicology, and paleoclimatology. For paleolimnologists, this could even lead to a better understanding of diagenesis and the distribution of organic molecules in ecosystems as transfer functions. The addition of new biomarkers would facilitate multi-proxy reconstructions, which would make historical reconstructions more detailed and paleolimnological research an even better resource for paleoclimate models to understand how climate change will affect us in the future.

Acknowledgements

This research was supported by a Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN-2018-04248) to JMB.

References

Alizadeh, B., Alipour, M., Chehrazi, A., Mirzaie, S., 2017. Chemometric classification and geochemistry of oils in the Iranian sector of the southern Persian Gulf Basin. Org. Geochem. 111, 67–81. https://doi.org/10.1016/j.orggeochem.2017.05.006.

Alshehry, Z.H., Barlow, C.K., Weir, J.M., Zhou, Y., McConville, M.J., Meikle, P.J., 2015. An efficient single phase method for the extraction of plasma lipids. Metabolites 5, 389–403. https://doi.org/10.3390/metabo5020389.

Armitage, E.G., Godzein, J., Alonso-Herranz, V., Lopez-Gonzalvez, A., Barbas, C., 2015. Missing value imputation strategies for metabolomics data. Electrophoresis 36, 3050–3060.

Asuero, A.G., Sayago, A., González, A.G., 2006. The correlation coefficient: an overview. Crit. Rev. Anal. Chem. 36, 41–59. https://doi.org/10.1080/10408340500526766.

Bader, T., Schulz, W., Kümmerer, K., Winzenbacher, R., 2016. General strategies to increase the repeatability in non-target screening by liquid chromatography-high resolution mass spectrometry. Anal. Chim. Acta 935, 173–186. https://doi.org/10.1016/j.aca.2016.06.030.

Beale, D.J., Crosswell, J., Karpe, A.V., Ahmed, W., Williams, M., Morrison, P.D., Metcalfe, S., Staley, C., Sadowsky, M.J., Palombo, E.A., Steven, A.D.L., 2017. A multi-omics based ecological analysis of coastal marine sediments from Gladstone, in Australia's Central Queensland, and Heron Island, a nearby fringing platform reef. Sci. Total Environ. 609, 842–853. https://doi.org/10.1016/j.scitotenv.2017.07.184.

Beale, D.J., Crosswell, J., Karpe, A.V., Metcalfe, S.S., Morrison, P.D., Staley, C., Ahmed, W., Sadowsky, M.J., Palombo, E.A., Steven, A.D.L., 2018. Seasonal metabolic analysis of marine sediments collected from Moreton Bay in South East Queensland, Australia, using a multi-omics-based approach. Sci. Total Environ. 631–632, 1328–1341. https://doi.org/10.1016/j.scitotenv.2018.03.106.

Berk, M., Ebbels, T., Montana, G., 2011. A statistical framework for biomarker discovery in metabolomic time course data. Bioinformatics 27, 1979–1985. https://doi.org/10.1093/bioinformatics/btr289.

Binder, H., Wirth, H., 2015. Analysis of large-scale OMIC data using self-organizing maps. Encyclopedia of Information Science and Technology. IGI Global, Hershey, PA, pp. 1642–1653.

Bingol, K., Zhang, F., Bruschweiler-Li, L., Brüschweiler, R., 2012. TOCCATA: a customized carbon total correlation spectroscopy NMR metabolomics database. Anal. Chem. 84, 9395–9401. https://doi.org/10.1021/ac302197e.

Bingol, K., Li, D.-W., Bruschweiler-Li, L., Cabrera, O.A., Megraw, T., Zhang, F., Brüschweiler, R., 2015. Unified and isomer-specific NMR metabolomics database for the accurate analysis of ^{13}C–^{1}H HSQC spectra. ACS Chem. Biol. 10, 452–459. https://doi.org/10.1021/cb5006382.

Blaise, B.J., Correia, G., Tin, A., Young, J.H., Vergnaud, A.-C., Lewis, M., Pearce, J.T.M., Elliott, P., Nicholson, J.K., Holmes, E., Ebbels, T.M.D., 2016. Power analysis and sample size determination in metabolic phenotyping. Anal. Chem. 88, 5179–5188. https://doi.org/10.1021/acs.analchem.6b00188.

Bligh, E.G., Dyer, W.J., 1959. A rapid method of total lipid extraction and purification. Can. J. Biochem. Physiol. 37, 911–917.

Böcker, S., Letzel, M.C., Lipták, Z., Pervukhin, A., 2009. SIRIUS: decomposing isotope patterns for metabolite identification†. Bioinformatics 25, 218–224. https://doi.org/10.1093/bioinformatics/btn603.

Brasili, E., Praticò, G., Marini, F., Valletta, A., Capuani, G., Sciubba, F., Miccheli, A., Pasqua, G., 2014. A non-targeted metabolomics approach to evaluate the effects of biomass growth and chitosan elicitation on primary and secondary metabolism of Hypericum perforatum in vitro roots. Metabolomics 10, 1186–1196. https://doi.org/10.1007/s11306-014-0660-z.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brereton, R.G., Lloyd, G.R., 2014. Partial least squares discriminant analysis: taking the magic away. J. Chemom. 28, 213–225. https://doi.org/10.1002/cem.2609.

Bu, Q., Wang, D., Liu, X., Wang, Z., 2014. A high throughout semi-quantification method for screening organic contaminants in river sediments. J. Environ. Manag. 143, 135–139. https://doi.org/10.1016/j.jenvman.2014.05.009.

Buszewski, B., Noga, S., 2012. Hydrophilic interaction liquid chromatography (HILIC)—a powerful separation technique. Anal. Bioanal. Chem. 402, 231–247. https://doi.org/10.1007/s00216-011-5308-5.

Cai, S.-S., Syage, J.A., 2006. Comparison of atmospheric pressure photoionization, atmospheric pressure chemical ionization, and electrospray ionization mass spectrometry for analysis of lipids. Anal. Chem. 78, 1191–1199. https://doi.org/10.1021/ac0515834.

Cajka, T., Fiehn, O., 2014. Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry. TrAC Trends Anal. Chem. 61, 192–206. https://doi.org/10.1016/j.trac.2014.04.017.

Cajka, T., Fiehn, O., 2016. Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. Anal. Chem. 88, 524–545. https://doi.org/10.1021/acs.analchem.5b04491.

Cambiaghi, A., Ferrario, M., Masseroli, M., 2016. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. Brief. Bioinform., bbw031 https://doi.org/10.1093/bib/bbw031.

Casilli, A., Silva, R.C., Laakia, J., Oliveira, C.J.F., Ferreira, A.A., Loureiro, M.R.B., Azevedo, D.A., Aquino Neto, F.R., 2014. High resolution molecular organic geochemistry assessment of Brazilian lacustrine crude oils. Org. Geochem. 68, 61–70. https://doi.org/10.1016/j.orggeochem.2014.01.009.

Cequier-Sánchez, E., Rodríguez, C., Ravelo, Á.G., Zárate, R., 2008. Dichloromethane as a solvent for lipid extraction and assessment of lipid classes and fatty acids from samples of different natures. J. Agric. Food Chem. 56, 4297–4303. https://doi.org/10.1021/jf073471e.

Chen, S., Hoene, M., Li, J., Li, Y., Zhao, X., Häring, H.-U., Schleicher, E.D., Weigert, C., Xu, G., Lehmann, R., 2013. Simultaneous extraction of metabolome and lipidome with methyl tert-butyl ether from a single small tissue sample for ultra-high performance liquid chromatography/mass spectrometry. J. Chromatogr. A 1298, 9–16. https://doi.org/10.1016/j.chroma.2013.05.019.

Chen, Y., Xu, J., Zhang, R., Shen, G., Song, Y., Sun, J., He, J., Zhan, Q., Abliz, Z., 2013. Assessment of data pre-processing methods for LC-MS/MS-based metabolomics of uterine cervix cancer. Analyst 138, 2669–2677. https://doi.org/10.1039/C3AN36818A.

Chen, J., Zhang, P., Lv, M., Guo, H., Huang, Y., Zhang, Z., Xu, F., 2017. Influences of normalization method on biomarker discovery in gas chromatography–mass spectrometry-based untargeted metabolomics: what should be considered? Anal. Chem. 89, 5342–5348.

Cherta, L., Portolés, T., Pitarch, E., Beltran, J., López, F.J., Calatayud, C., Company, B., Hernández, F., 2015. Analytical strategy based on the combination of gas chromatography coupled to time-of-flight and hybrid quadrupole time-of-flight mass analyzers for non-target analysis in food packaging. Food Chem. 188, 301–308. https://doi.org/10.1016/j.foodchem.2015.04.141.

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D.S., Xia, J., 2018. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res. https://doi.org/10.1093/nar/gky310.

Claesen, J., Dittwald, P., Burzykowski, T., Valkenborg, D., 2012. An efficient method to calculate the aggregated isotopic distribution and exact center-masses. J. Am. Soc. Mass Spectrom. 23, 753–763. https://doi.org/10.1007/s13361-011-0326-2.

Coble, J.B., Fraga, C.G., 2014. Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery. J. Chromatogr. A 1358, 155–164. https://doi.org/10.1016/j.chroma.2014.06.100.

Considine, E.C., Thomas, G., Boulesteix, A.L., Khashan, A.S., Kenny, L.C., 2018. Critical review of reporting of the data analysis step in metabolomics. Metabolomics 14, 7. https://doi.org/10.1007/s11306-017-1299-3.

Creek, D.J., Dunn, W.B., Fiehn, O., Griffin, J.L., Hall, R.D., Lei, Z., Mistrik, R., Neumann, S., Schymanski, E.L., Sumner, L.W., Trengove, R., Wolfender, J.-L., 2014. Metabolite identification: are you sure? And how do your peers gauge your confidence? Metabolomics 10, 350–353. https://doi.org/10.1007/s11306-014-0656-8.

Cui, Q., Lewis, I.A., Hegeman, A.D., Anderson, M.E., Li, J., Schulte, C.F., Westler, W.M., Eghbalnia, H.R., Sussman, M.R., Markley, J.L., 2008. Metabolite identification via the Madison Metabolomics Consortium Database. Nat. Biotechnol. 26, 162–164. https://doi.org/10.1038/nbt0208-162.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792. https://doi.org/10.1890/07-0539.1.

D'Andrilli, J., Foreman, C.M., Marshall, A.G., McKnight, D.M., 2013. Characterization of IHSS Pony Lake fulvic acid dissolved organic matter by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry and fluorescence spectroscopy. Org. Geochem. 65, 19–28. https://doi.org/10.1016/j.orggeochem.2013.09.013.

Davidson, R.L., Weber, R.J.M., Liu, H., Sharma-Oates, A., Viant, M.R., 2016. Galaxy-M: a galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. GigaScience 5. https://doi.org/10.1186/s13742-016-0115-8.

De Livera, A.M., Dias, D.A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., Roessner, U., McConville, M., Speed, T.P., 2012. Normalizing and integrating metabolomics data. Anal. Chem. 84, 10768–10776.

Di Guida, R., Engel, J., Allwood, J.W., Weber, R.J.M., Jones, M.R., Sommer, U., Viant, M.R., Dunn, W.B., 2016. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics 12. https://doi.org/10.1007/s11306-016-1030-9.

Dudzik, D., Barbas-Bernardos, C., García, A., Barbas, C., 2018. Quality assurance procedures for mass spectrometry untargeted metabolomics. a review. J. Pharm. Biomed. Anal. 147, 149–173. https://doi.org/10.1016/j.jpba.2017.07.044.

Dunn, W.B., Wilson, I.D., Nicholls, A.W., Broadhurst, D., 2012. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. Bioanalysis 4, 2249–2264. https://doi.org/10.4155/bio.12.204.

Dunn, W.B., Erban, A., Weber, R.J.M., Creek, D.J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J., Viant, M.R., 2013. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics 9, 44–66. https://doi.org/10.1007/s11306-012-0434-4.

Dwivedi, P., Schultz, A.J., Jr, H.H.H., 2010. Metabolic profiling of human blood by high-resolution ion mobility mass spectrometry (IM-MS). Int. J. Mass Spectrom. 298, 78–90. https://doi.org/10.1016/j.ijms.2010.02.007.

Fahy, E., Sud, M., Cotter, D., Subramaniam, S., 2007. LIPID MAPS online tools for lipid research. Nucleic Acids Res. 35, W606–W612. https://doi.org/10.1093/nar/gkm324.

Fahy, E., Alvarez-Jarreta, J., Brasher, C.J., Nguyen, A., Hawksworth, J.I., Rodrigues, P., Meckelmann, S., Allen, S.M., O'Donnell, V.B., 2018. LipidFinder on LIPID MAPS: peak filtering, MS searching and statistical analysis for lipidomics. Bioinformatics https://doi.org/10.1093/bioinformatics/bty679.

Farrés, M., Martrat, B., de Mol, B., Grimalt, J.O., Tauler, R., 2015a. Extraction of climatic signals from fossil organic compounds in marine sediments up to 11.7Ma old (IODP-U1318). Anal. Chim. Acta 879, 1–9. https://doi.org/10.1016/j.aca.2015.04.051.

Farrés, M., Platikanov, S., Tsakovski, S., Tauler, R., 2015b. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. J. Chemom. 29, 528–536.

Fernández-Albert, F., Llorach, R., Andrés-Lacueva, C., Perera, A., 2014. An R package to analyse LC/MS metabolomic data: MAIT (Metabolite Automatic Identification Toolkit). Bioinformatics 30, 1937–1939. https://doi.org/10.1093/bioinformatics/btu136.

Fernandez-de-Cossio Diaz, J., Fernandez-de-Cossio, J., 2012. Computation of isotopic peak center-mass distribution by Fourier Transform. Anal. Chem. 84, 7052–7056. https://doi.org/10.1021/ac301296a.

Fernandez-Lima, F.A., Becker, C., McKenna, A.M., Rodgers, R.P., Marshall, A.G., Russell, D.H., 2009. Petroleum crude oil characterization by IMS-MS and FTICR MS. Anal. Chem. 81, 9941–9947. https://doi.org/10.1021/ac901594f.

Fiehn, O., Robertson, D., Griffin, J., van der Werf, M., Nikolau, B., Morrison, N., Sumner, L.W., Goodacre, R., Hardy, N.W., Taylor, C., Fostel, J., Kristal, B., Kaddurah-Daouk, R., Mendes, P., van Ommen, B., Lindon, J.C., Sansone, S.-A., 2007. The metabolomics standards initiative (MSI). Metabolomics 3, 175–178. https://doi.org/10.1007/s11306-007-0070-6.

Folch, J., Lees, M., Stanley, G.H.S., 1956. A simple method for the isolation and purification of total lipides from animal tissues. J. Biol. Chem. 226, 497–509.

Forsberg, E.M., Huan, T., Rinehart, D., Benton, H.P., Warth, B., Hilmers, B., Siuzdak, G., 2018. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. Nat. Protoc. 13, 633–651. https://doi.org/10.1038/nprot.2017.151.

Franceschi, P., Wehrens, R., 2014. Self-organizing maps: a versatile tool for the automatic analysis of untargeted imaging datasets. Proteomics 14, 853–861.

Furey, A., Moriarty, M., Bane, V., Kinsella, B., Lehane, M., 2013. Ion suppression; a critical review on causes, evaluation, prevention and applications. Talanta 115, 104–122. https://doi.org/10.1016/j.talanta.2013.03.048.

Fushimi, A., Hashimoto, S., Ieda, T., Ochiai, N., Takazawa, Y., Fujitani, Y., Tanabe, K., 2012. Thermal desorption – comprehensive two-dimensional gas chromatography coupled with tandem mass spectrometry for determination of trace polycyclic aromatic hydrocarbons and their derivatives. J. Chromatogr. A 1252, 164–170. https://doi.org/10.1016/j.chroma.2012.06.068.

Gathungu, R.M., Kautz, R., Kristal, B.S., Bird, S.S., Vouros, P., 2018. The integration of LC-MS and NMR for the analysis of low molecular weight trace analytes in complex matrices. Mass Spectrom. Rev., 1–19 https://doi.org/10.1002/mas.21575.

Giacomoni, F., Le Corguillé, G., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., Duperier, C., Tremblay-Franco, M., Martin, J.-F., Jacob, D., Goulitquer, S., Thévenot, E.A., Caron, C., 2015. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. Bioinformatics 31, 1493–1495. https://doi.org/10.1093/bioinformatics/btu813.

Giancarlo, R., Lo Bosco, G., Pinello, L., 2010. Distance functions, clustering algorithms and microarray data analysis. In: Blum, C., Battiti, R. (Eds.), Learning and Intelligent Optimization. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 125–138 https://doi.org/10.1007/978-3-642-13800-3_10.

Giancarlo, R., Bosco, G.L., Pinello, L., Utro, F., 2011. The three steps of clustering in the post-genomic era: a synopsis. In: Rizzo, R., Lisboa, P.J.G. (Eds.), Computational Intelligence Methods for Bioinformatics and Biostatistics. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 13–30 https://doi.org/10.1007/978-3-642-21946-7_2.

Gil de la Fuente, A., Grace Armitage, E., Otero, A., Barbas, C., Godzien, J., 2017. Differentiating signals to make biological sense - a guide through databases for MS-based non-targeted metabolomics: general. ELECTROPHORESIS 38, 2242–2256. https://doi.org/10.1002/elps.201700070.

Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G., Kell, D.B., 2004. Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol. 22, 245–252. https://doi.org/10.1016/j.tibtech.2004.03.007.

Gorrochategui, E., Jaumot, J., Lacorte, S., Tauler, R., 2016. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow. Trends Anal. Chem. 82, 425–442. https://doi.org/10.1016/j.trac.2016.07.004.

Gregory, K.E., Bird, S.S., Gross, V.S., Marur, V.R., Lazarev, A.V., Walker, W.A., Kristal, B.S., 2012. Method development for fecal lipidomics profiling. Anal. Chem. 85, 1114–1123.

Grigoriadou, A., Schwarzbauer, J., 2011. Non-target screening of organic contaminants in sediments from the industrial coastal area of Kavala City (NE Greece). Water Air Soil Pollut. 214, 623–643. https://doi.org/10.1007/s11270-010-0451-8.

Gromski, P.S., Xu, Y., Correa, E., Ellis, D.I., Turner, M.L., Goodacre, R., 2014a. A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. Anal. Chim. Acta 829, 1–8. https://doi.org/10.1016/j.aca.2014.03.039.

Gromski, P.S., Xu, Y., Kotze, H.L., Correa, E., Ellis, D.I., Armitage, E.G., Turner, M.L., Goodacre, R., 2014b. Influence of missing values substitutes on multivariate analysis of metabolomics data. Metabolites 4, 433–452. https://doi.org/10.3390/metabo4020433.

Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L., Goodacre, R., 2015. A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. Anal. Chim. Acta 879, 10–23. https://doi.org/10.1016/j.aca.2015.02.012.

Gros, J., Reddy, C.M., Aeppli, C., Carmichael, C.A., Arey, J.S., 2014. Resolving biodegradation patterns of persistent saturated hydrocarbons in weathered oil samples from the *Deepwater Horizon* disaster. Environ. Sci. Technol. 48, 1628–1637. https://doi.org/10.1021/es4042836.

Gruber, B., Weggler, B.A., Jaramillo, R., Murrell, K.A., Piotrowski, P.K., Dorman, F.L., 2018. Comprehensive two-dimensional gas chromatography in forensic science: a critical review of recent trends. TrAC Trends Anal. Chem. https://doi.org/10.1016/j.trac.2018.05.017.

Guijas, C., Montenegro-Burke, J.R., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., Koellensperger, G., Huan, T., Uritboonthai, W., Aisporna, A.E., Wolan, D.W., Spilker, M.E., Benton, H.P., Siuzdak, G., 2018. METLIN: a technology platform for identifying knowns and unknowns. Anal. Chem. 90, 3156–3164. https://doi.org/10.1021/acs.analchem.7b04424.

Gürdeniz, G., Kristensen, M., Skov, T., Dragsted, L.O., 2012. The effect of LC-MS data preprocessing methods on the selection of plasma biomarkers in fed vs. fasted rats. Metabolites 2, 77–99. https://doi.org/10.3390/metabo2010077.

Gustavsson, J., Wiberg, K., Ribeli, E., Nguyen, M.A., Josefsson, S., Ahrens, L., 2018. Screening of organic flame retardants in Swedish river water. Sci. Total Environ. 625, 1046–1055. https://doi.org/10.1016/j.scitotenv.2017.12.281.

Gutiérrez Sama, S., Farenc, M., Barrère-Mangote, C., Lobinski, R., Afonso, C., Bouyssière, B., Giusti, P., 2018. Molecular fingerprints and speciation of crude oils and heavy fractions revealed by molecular and elemental mass spectrometry: keystone between petroleomics, metallopetroleomics, and petrointeractomics. Energy Fuel 32, 4593–4605. https://doi.org/10.1021/acs.energyfuels.7b03218.

Haddad, I., Hiller, K., Frimmersdorf, E., Benkert, B., Schomburg, D., Jahn, D., 2009. An emergent self-organizing map based analysis pipeline for comparative metabolome studies. In Silico Biol. 9, 163–178. https://doi.org/10.3233/ISB-2009-0396.

Hall, G.J., Frysinger, G.S., Aeppli, C., Carmichael, C.A., Gros, J., Lemkau, K.L., Nelson, R.K., Reddy, C.M., 2013. Oxygenated weathering products of Deepwater Horizon oil come from surprising precursors. Mar. Pollut. Bull. 75, 140–149. https://doi.org/10.1016/j.marpolbul.2013.07.048.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Unsupervised learning. The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY, pp. 485–585 https://doi.org/10.1007/978-0-387-84858-7_14.

Haun, J., Leonhardt, J., Portner, C., Hetzel, T., Tuerk, J., Teutenberg, T., Schmidt, T.C., 2013. Online and splitless NanoLC × CapillaryLC with quadrupole/time-of-flight mass spectrometric detection for comprehensive screening analysis of complex samples. Anal. Chem. 85, 10083–10090. https://doi.org/10.1021/ac402002m.

Hernández, F., Sancho, J.V., Ibáñez, M., Abad, E., Portolés, T., Mattioli, L., 2012. Current use of high-resolution mass spectrometry in the environmental sciences. Anal. Bioanal. Chem. 403, 1251–1264. https://doi.org/10.1007/s00216-012-5844-7.

Heude, C., Nath, J., Carrigan, J.B., Ludwig, C., 2017. Nuclear magnetic resonance strategies for metabolic analysis. Metabolomics: From Fundamentals to Clinical Applications. Advances in Experimental Medicine and Biology. Springer, Cham, pp. 45–76 https://doi.org/10.1007/978-3-319-47656-8_3.

Hilton, D.C., Jones, R.S., Sjödin, A., 2010. A method for rapid, non-targeted screening for environmental contaminants in household dust. J. Chromatogr. A 1217, 6851–6856. https://doi.org/10.1016/j.chroma.2010.08.039.

Hinz, C., Liggi, S., Griffin, J.L., 2018. The potential of ion mobility mass spectrometry for high-throughput and high-resolution lipidomics. Curr. Opin. Chem. Biol. 42, 42–50. https://doi.org/10.1016/j.cbpa.2017.10.018.

Holčapek, M., Ovčačíková, M., Lísa, M., Cífková, E., Hájek, T., 2015. Continuous comprehensive two-dimensional liquid chromatography–electrospray ionization mass spectrometry of complex lipidomic samples. Anal. Bioanal. Chem. 407, 5033–5043. https://doi.org/10.1007/s00216-015-8528-2.

Hollender, J., Schymanski, E.L., Singer, H.P., Ferguson, P.L., 2017. Nontarget screening with high resolution mass spectrometry in the environment: ready to go? Environ. Sci. Technol. 51, 11505–11512. https://doi.org/10.1021/acs.est.7b02184.

Holman, J.D., Tabb, D.L., Mallick, P., 2015. Employing ProteoWizard to convert raw mass spectrometry data. Curr. Protoc. Bioinformatics 46, 2–13.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, Kenichi, Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M.Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, Ken, Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., Nishioka, T., 2010. MassBank: a public repository for sharing mass spectral data for life sciences. J. Mass Spectrom. 45, 703–714. https://doi.org/10.1002/jms.1777.

Hur, M., Kim, S., Hsu, C.S., 2017. Petroinformatics, in: Springer Handbook of Petroleum Technology. Springer Handbooks. Springer, Cham, pp. 173–198 https://doi.org/10.1007/978-3-319-49347-3_4.

Hur, M., Ware, R.L., Park, J., McKenna, A.M., Rodgers, R.P., Nikolau, B.J., Wurtele, E.S., Marshall, A.G., 2018. Statistically significant differences in composition of petroleum crude oils revealed by volcano plots generated from ultrahigh resolution Fourier transform ion cyclotron resonance mass spectra. Energy Fuel 32, 1206–1212. https://doi.org/10.1021/acs.energyfuels.7b03061.

Ieda, T., Ochiai, N., Miyawaki, T., Ohura, T., Horii, Y., 2011. Environmental analysis of chlorinated and brominated polycyclic aromatic hydrocarbons by comprehensive two-dimensional gas chromatography coupled to high-resolution time-of-flight mass spectrometry. J. Chromatogr. A 1218, 3224–3232. https://doi.org/10.1016/j.chroma.2011.01.013.

Imbert, L., Gaudin, M., Libong, D., Touboul, D., Abreu, S., Loiseau, P.M., Laprévote, O., Chaminade, P., 2012. Comparison of electrospray ionization, atmospheric pressure chemical ionization and atmospheric pressure photoionization for a lipidomic analysis of *Leishmania donovani*. J. Chromatogr. A 1242, 75–83. https://doi.org/10.1016/j.chroma.2012.04.035.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. 31, 264–323. https://doi.org/10.1145/331499.331504.

Jaskowiak, P.A., Campello, R.J.G.B., Costa, I.G., 2013. Proximity measures for clustering gene expression microarray data: a validation methodology and a comparative analysis. IEEE/ACM Trans. Comput. Biol. Bioinform. 10, 845–857. https://doi.org/10.1109/TCBB.2013.9.

Jaskowiak, P.A., Campello, R.J., Costa, I.G., 2014. On the selection of appropriate distances for gene expression data clustering. BMC Bioinf. 15 (S2). https://doi.org/10.1186/1471-2105-15-S2-S2.

Jiménez-Morillo, N.T., González-Pérez, J.A., Almendros, G., De la Rosa, J.M., Waggoner, D.C., Jordán, A., Zavala, L.M., González-Vila, F.J., Hatcher, P.G., 2018. Ultra-high resolution mass spectrometry of physical speciation patterns of organic matter in fire-affected soils. J. Environ. Manag. 225, 139–147. https://doi.org/10.1016/j.jenvman.2018.07.069.

Jurowski, K., Kochan, K., Walczak, J., Barańska, M., Piekoszewski, W., Buszewski, B., 2017. Comprehensive review of trends and analytical strategies applied for biological samples preparation and storage in modern medical lipidomics: state of the art. TrAC Trends Anal. Chem. 86, 276–289. https://doi.org/10.1016/j.trac.2016.10.014.

Kamínski, M., Kartanowicz, R., Gilgenast, E., Namieśnik, J., 2005. High-performance liquid chromatography in group-type separation and technical or process analytics of petroleum products. Crit. Rev. Anal. Chem. 35, 193–216. https://doi.org/10.1080/10408340500304024.

Karaman, I., 2017. Preprocessing and pretreatment of metabolomics data for statistical analysis. Metabolomics: From Fundamentals to Clinical Applications, Advances in Experimental Medicine and Biology. Springer, Cham, pp. 145–161 https://doi.org/10.1007/978-3-319-47656-8_6.

Kasprzak, E.M., Lewis, K.E., 2001. Pareto analysis in multiobjective optimization using the collinearity theorem and scaling method. Struct. Multidiscip. Optim. 22, 208–218. https://doi.org/10.1007/s001580100138.

Kiepper, A.P., Casilli, A., Azevedo, D.A., 2014. Depositional paleoenvironment of Brazilian crude oils from unusual biomarkers revealed using comprehensive two dimensional gas chromatography coupled to time of flight mass spectrometry. Org. Geochem. 70, 62–75. https://doi.org/10.1016/j.orggeochem.2014.03.005.
Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., Wang, J., Yu, B., Zhang, J., Bryant, S.H., 2016. PubChem substance and compound databases. Nucleic Acids Res. 44, D1202–D1213. https://doi.org/10.1093/nar/gkv951.
Kind, T., Fiehn, O., 2006. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. BMC Bioinf. 7, 234. https://doi.org/10.1186/1471-2105-7-234.
Kind, T., Fiehn, O., 2007. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinf. 8, 105. https://doi.org/10.1186/1471-2105-8-105.
Kirpich, A.S., Ibarra, M., Moskalenko, O., Fear, J.M., Gerken, J., Mi, X., Ashrafi, A., Morse, A.M., McIntyre, L.M., 2018. SECIMTools: a suite of metabolomics data analysis tools. BMC Bioinf. 19. https://doi.org/10.1186/s12859-018-2134-1.
Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmuller, E., Dormann, P., Weckwerth, W., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A.R., Steinhauser, D., 2005. GMD@CSB.DB: the Golm metabolome database. Bioinformatics 21, 1635–1638. https://doi.org/10.1093/bioinformatics/bti236.
Krauss, M., Singer, H., Hollender, J., 2010. LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns. Anal. Bioanal. Chem. 397, 943–951. https://doi.org/10.1007/s00216-010-3608-9.
van Krevelen, D.W., 1950. Graphical-statistical method for the study of structure and reaction processes of coal. Fuel 29, 269–284.
Kronimus, A., Schwarzbauer, J., 2007. Non-target screening of extractable and non-extractable organic xenobiotics in riverine sediments of Ems and Mulde Rivers, Germany. Environ. Pollut. 147, 176–186. https://doi.org/10.1016/j.envpol.2006.08.014.
Krzywinski, M., Altman, N., 2013. Power and sample size: points of significance. Nat. Methods 10, 1139–1140. https://doi.org/10.1038/nmeth.2738.
Kujawinski, E.B., 2011. The impact of microbial metabolism on marine dissolved organic matter. Annu. Rev. Mar. Sci. 3, 567–599. https://doi.org/10.1146/annurev-marine-120308-081003.
Kuznetsova, N., 2014. Random Forest Visualization. Eindhoven University of Technology.
Kvalheim, O.M., 1985. Scaling of analytical data. Anal. Chim. Acta 177, 71–79.
Laakia, J., Casilli, A., Araújo, B.Q., Gonçalves, F.T.T., Marotta, E., Oliveira, C.J.F., Carbonezi, C.A., Loureiro, M.R.B., Azevedo, D.A., Aquino Neto, F.R., 2017. Characterization of unusual tetracyclic compounds and possible novel maturity parameters for Brazilian crude oils using comprehensive two-dimensional gas chromatography-time of flight mass spectrometry. Org. Geochem. 106, 93–104. https://doi.org/10.1016/j.orggeochem.2016.10.012.
Lanucara, F., Holman, S.W., Gray, C.J., Eyers, C.E., 2014. The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. Nat. Chem. 6, 281–294. https://doi.org/10.1038/nchem.1889.
Lê Cao, K.-A., Boitard, S., Besse, P., 2011. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinf. 12, 253. https://doi.org/10.1186/1471-2105-12-253.
Li, B., Tang, J., Yang, Q., Cui, X., Li, S., Chen, S., Cao, Q., Xue, W., Chen, N., Zhu, F., 2016. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. Sci. Rep. 6, 1–13.
Li, J., Vosegaard, T., Guo, Z., 2017. Applications of nuclear magnetic resonance in lipid analyses: an emerging powerful tool for lipidomics studies. Prog. Lipid Res. 68, 37–56. https://doi.org/10.1016/j.plipres.2017.09.003.
Li, X.M., Sun, G.-X., Chen, S.-C., Fang, Z., Yuan, H.-Y., Shi, Q., Zhu, Y.-G., 2018. Molecular chemodiversity of dissolved organic matter in paddy soils. Environ. Sci. Technol. 52, 963–971. https://doi.org/10.1021/acs.est.7b00377.
Li, Z., Lu, Y., Guo, Y., Cao, H., Wang, Q., Shui, W., 2018. Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection. Anal. Chim. Acta 1029, 50–57.
Löfgren, L., Forsberg, G.-B., Ståhlman, M., 2016. The BUME method: a new rapid and simple chloroform-free method for total lipid extraction of animal tissue. Sci. Rep. 6, 27688. https://doi.org/10.1038/srep27688.
Mahadevan, S., Shah, S.L., Marrie, T.J., Slupsky, C.M., 2008. Analysis of metabolomic data using support vector machines. Anal. Chem. 80, 7562–7570. https://doi.org/10.1021/ac800954c.
Mairinger, T., Causon, T.J., Hann, S., 2018. The potential of ion mobility–mass spectrometry for non-targeted metabolomics. Curr. Opin. Chem. Biol. 42, 9–15. https://doi.org/10.1016/j.cbpa.2017.10.015.
Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W.H., Römpp, A., Neumann, S., Pizarro, A.D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., Deutsch, E.W., 2011. mzML—a community standard for mass spectrometry data. Mol. Cell. Proteomics 10, R110.000133. https://doi.org/10.1074/mcp.R110.000133.
Martins, M.C.M., Caldana, C., Wolf, L.D., Abreu, L.G.F. de, 2018. The importance of experimental design, quality assurance, and control in plant metabolomics experiments, in: Plant Metabolomics, Methods in Molecular Biology. Humana Press, New York, NY, pp. 3–17. doi:https://doi.org/10.1007/978-1-4939-7819-9_1.
Masoom, H., Courtier-Murias, D., Farooq, H., Soong, R., Kelleher, B.P., Zhang, C., Maas, W.E., Fey, M., Kumar, R., Monette, M., 2016. Soil organic matter in its native state: unravelling the most complex biomaterial on earth. Environ. Sci. Technol. 50, 1670–1680.
Matyash, V., Liebisch, G., Kurzchalia, T.V., Shevchenko, A., Schwudke, D., 2008. Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. J. Lipid Res. 49, 1137–1146.
Melamud, E., Vastag, L., Rabinowitz, J.D., 2010. Metabolomic analysis and visualization engine for LC−MS data. Anal. Chem. 82, 9818–9826. https://doi.org/10.1021/ac1021166.
Meyers, P.A., Ishiwatari, R., 1993. Lacustrine organic geochemistry-an overview of indicators of organic matter sources and diagenesis in lake sediments. Org. Geochem. 20, 867–900. https://doi.org/10.1016/0146-6380(93)90100-p.
Mondello, L., Tranchida, P.Q., Dugo, P., Dugo, G., 2008. Comprehensive two-dimensional gas chromatography-mass spectrometry: a review. Mass Spectrom. Rev. 27, 101–124. https://doi.org/10.1002/mas.20158.
Mopper, K., Stubbins, A., Ritchie, J.D., Bialk, H.M., Hatcher, P.G., 2007. Advanced instrumental approaches for characterization of marine dissolved organic matter: extraction techniques, mass spectrometry, and nuclear magnetic resonance spectroscopy. Chem. Rev. 107, 419–442. https://doi.org/10.1021/cr050359b.
Morris, A.D., Letcher, R.J., Dyck, M., Chandramouli, B., Fisk, A.T., Cosgrove, J., 2018. Multivariate statistical analysis of metabolomics profiles in tissues of polar bears (*Ursus maritimus*) from the Southern and Western Hudson Bay subpopulations. Polar Biol. 41, 433–449.
Murray, K.K., Boyd, R.K., Eberlin, M.N., Langley, G.J., Li, L., Naito, Y., 2013. Definitions of terms relating to mass spectrometry (IUPAC recommendations 2013). Pure Appl. Chem. 85, 1515–1609. https://doi.org/10.1351/PAC-REC-06-04-06.
Myers, O.D., Sumner, S.J., Li, S., Barnes, S., Du, X., 2017a. Detailed investigation and comparison of the XCMS and MZmine 2 chromatogram construction and chromatographic peak detection methods for preprocessing mass spectrometry metabolomics data. Anal. Chem. 89, 8689–8695. https://doi.org/10.1021/acs.analchem.7b01069.
Myers, O.D., Sumner, S.J., Li, S., Barnes, S., Du, X., 2017b. One step forward for reducing false positive and false negative compound identifications from mass spectrometry metabolomics data: new algorithms for constructing extracted ion chromatograms and detecting chromatographic peaks. Anal. Chem. 89, 8696–8703. https://doi.org/10.1021/acs.analchem.7b00947.
Navarro-Reig, M., Jaumot, J., Tauler, R., 2018. An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis. J. Chromatogr. A https://doi.org/10.1016/j.chroma.2018.07.017.
Nizio, K.D., McGinitie, T.M., Harynuk, J.J., 2012. Comprehensive multidimensional separations for the analysis of petroleum. J. Chromatogr. A 1255, 12–23. https://doi.org/10.1016/j.chroma.2012.01.078.
Nováková, L., Vlčková, H., 2009. A review of current trends and advances in modern bioanalytical methods: chromatography and sample preparation. Anal. Chim. Acta 656, 8–35. https://doi.org/10.1016/j.aca.2009.10.004.
Orchard, S., Montechi-Palazzi, L., Deutsch, E.W., Binz, P.-A., Jones, A.R., Paton, N., Pizarro, A., Creasy, D.M., Wojcik, J., Hermjakob, H., 2007. Five years of progress in the standardization of proteomics data 4th annual spring workshop of the HUPO-proteomics standards initiative April 23–25, 2007 Ecole Nationale Supérieure (ENS), Lyon, France. PROTEOMICS 7, 3436–3440. https://doi.org/10.1002/pmic.200700658.
Ouyang, X., Leonards, P., Legler, J., van der Oost, R., de Boer, J., Lamoree, M., 2015. Comprehensive two-dimensional liquid chromatography coupled to high resolution time of flight mass spectrometry for chemical characterization of sewage treatment plant effluents. J. Chromatogr. A 1380, 139–145. https://doi.org/10.1016/j.chroma.2014.12.075.
Ouyang, X., Weiss, J.M., de Boer, J., Lamoree, M.H., Leonards, P.E.G., 2017. Non-target analysis of household dust and laundry dryer lint using comprehensive two-dimensional liquid chromatography coupled with time-of-flight mass spectrometry. Chemosphere 166, 431–437. https://doi.org/10.1016/j.chemosphere.2016.09.107.
Paglia, G., Astarita, G., 2017. Metabolomics and lipidomics using traveling-wave ion mobility mass spectrometry. Nat. Protoc. 12, 797–813. https://doi.org/10.1038/nprot.2017.013.
Paglia, G., Williams, J.P., Menikarachchi, L., Thompson, J.W., Tyldesley-Worster, R., Halldórsson, S., Rolfsson, O., Moseley, A., Grant, D., Langridge, J., Palsson, B.O., Astarita, G., 2014. Ion mobility derived collision cross sections to support metabolomics applications. Anal. Chem. 86, 3985–3993. https://doi.org/10.1021/ac500405x.
Paglia, G., Angel, P., Williams, J.P., Richardson, K., Olivos, H.J., Thompson, J.W., Menikarachchi, L., Lai, S., Walsh, C., Moseley, A., Plumb, R.S., Grant, D.F., Palsson, B.O., Langridge, J., Geromanos, S., Astarita, G., 2015. Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification. Anal. Chem. 87, 1137–1144. https://doi.org/10.1021/ac503715v.
Pani, O., Górecki, T., 2006. Comprehensive two-dimensional gas chromatography (GC×GC) in environmental analysis and monitoring. Anal. Bioanal. Chem. 386, 1013–1023. https://doi.org/10.1007/s00216-006-0568-1.
Parastar, H., Garreta-Lara, E., Campos, B., Barata, C., Lacorte, S., Tauler, R., 2018. Chemometrics comparison of gas chromatography with mass spectrometry and comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry *Daphnia magna* metabolic profiles exposed to salinity. J. Sep. Sci. 41, 2368–2379. https://doi.org/10.1002/jssc.201701336.
Pena-Abaurrea, M., Jobst, K.J., Ruffolo, R., Shen, L., McCrindle, R., Helm, P.A., Reiner, E.J., 2014. Identification of potential novel bioaccumulative and persistent chemicals in sediments from Ontario (Canada) using scripting approaches with GC×GC-TOF MS analysis. Environ. Sci. Technol. 48, 9591–9599. https://doi.org/10.1021/es5018152.
Pence, H.E., Williams, A., 2010. ChemSpider: an online chemical information resource. J. Chem. Educ. 87, 1123–1124. https://doi.org/10.1021/ed100697w.
Pinto, R.C., 2017. Chemometrics methods and strategies in metabolomics. Metabolomics: From Fundamentals to Clinical Applications, Advances in Experimental Medicine and Biology. Springer, Cham, pp. 163–190 https://doi.org/10.1007/978-3-319-47656-8_7.
Pisani, O., Haddix, M.L., Conant, R.T., Paul, E.A., Simpson, M.J., 2016. Molecular composition of soil organic matter with land-use change along a bi-continental mean annual temperature gradient. Sci. Total Environ. 573, 470–480.

Pluskal, T., Castillo, S., Villar-Briones, A., Orešič, M., 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinf. 11, 395. https://doi.org/10.1186/1471-2105-11-395.

Poad, B.L.J., Zheng, X., Mitchell, T.W., Smith, R.D., Baker, E.S., Blanksby, S.J., 2018. Online ozonolysis combined with ion mobility-mass spectrometry provides a new platform for lipid isomer analyses. Anal. Chem. 90, 1292–1300. https://doi.org/10.1021/acs.analchem.7b04091.

Pollo, B.J., Alexandrino, G.L., Augusto, F., Hantao, L.W., 2018. The impact of comprehensive two-dimensional gas chromatography on oil & gas analysis: recent advances and applications in petroleum industry. TrAC Trends Anal. Chem. 105, 202–217. https://doi.org/10.1016/j.trac.2018.05.007.

Prata, P.S., Alexandrino, G.L., Mogollón, N.G.S., Augusto, F., 2016. Discriminating Brazilian crude oils using comprehensive two-dimensional gas chromatography–mass spectrometry and multiway principal component analysis. J. Chromatogr. A 1472, 99–106. https://doi.org/10.1016/j.chroma.2016.10.044.

Raterink, R.-J., Lindenburg, P.W., Vreeken, R.J., Ramautar, R., Hankemeier, T., 2014. Recent developments in sample-pretreatment techniques for mass spectrometry-based metabolomics. TrAC Trends Anal. Chem. 61, 157–167. https://doi.org/10.1016/j.trac.2014.06.003.

Reis, A., Rudnitskaya, A., Blackburn, G.J., Fauzi, N.M., Pitt, A.R., Spickett, C.M., 2013. A comparison of five lipid extraction solvent systems for lipidomic studies of human LDL. J. Lipid Res. 54, 1812–1824. https://doi.org/10.1194/jlr.M034330.

Ren, S., Hinzman, A.A., Kang, E.L., Szczesniak, R.D., Lu, L.J., 2015. Computational and statistical analysis of metabolomics data. Metabolomics 11, 1492–1513. https://doi.org/10.1007/s11306-015-0823-6.

Rivas-Ubach, A., Liu, Y., Bianchi, T.S., Tolić, N., Jansson, C., Paša-Tolić, L., 2018. Moving beyond the van Krevelen Diagram: a new stoichiometric approach for compound classification in organisms. Anal. Chem. 90, 6152–6160. https://doi.org/10.1021/acs.analchem.8b00529.

Robertson, D.G., 2005. Metabonomics in toxicology: a review. Toxicol. Sci. 85, 809–822. https://doi.org/10.1093/toxsci/kfi102.

Rochat, B., 2017. Proposed confidence scale and ID score in the identification of known-unknown compounds using high resolution MS data. J. Am. Soc. Mass Spectrom. 28, 709–723. https://doi.org/10.1007/s13361-016-1556-0.

Rodgers, R.P., McKenna, A.M., 2011. Petroleum analysis. Anal. Chem. 83, 4665–4687. https://doi.org/10.1021/ac201080e.

Rostkowski, P., Haglund, P., Dye, C., Schlabach, M., 2013. Non-target screening of environmental samples by low and high resolution time of flight mass spectrometry (TOF-MS). Proc. 13th Int. Conf. Envrionmental Sci. Technol, pp. 1–4.

Rusilowicz, M., Dickinson, M., Charlton, A., O'Keefe, S., Wilson, J., 2016. A batch correction method for liquid chromatography–mass spectrometry data that does not depend on quality control samples. Metabolomics 12, 56. https://doi.org/10.1007/s11306-016-0972-2.

Ruttkies, C., Schymanski, E.L., Wolf, S., Hollender, J., Neumann, S., 2016. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. Aust. J. Chem. 8. https://doi.org/10.1186/s13321-016-0115-9.

Sauerschnig, C., Doppler, M., Bueschl, C., Schuhmacher, R., 2017. Methanol generates numerous artifacts during sample extraction and storage of extracts in metabolomics research. Metabolites 8 (1). https://doi.org/10.3390/metabo8010001.

Scheubert, K., Hufsky, F., Böcker, S., 2013. Computational mass spectrometry for small molecules. Aust. J. Chem. 5, 12. https://doi.org/10.1186/1758-2946-5-12.

Schymanski, E.L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H.P., Hollender, J., 2014. Identifying small molecules via high resolution mass spectrometry: communicating confidence. Environ. Sci. Technol. 48, 2097–2098. https://doi.org/10.1021/es5002105.

Serrano, R., Nácher-Mestre, J., Portolés, T., Amat, F., Hernández, F., 2011. Non-target screening of organic contaminants in marine salts by gas chromatography coupled to high-resolution time-of-flight mass spectrometry. Talanta 85, 877–884. https://doi.org/10.1016/j.talanta.2011.04.055.

Silva, R.S.F., Aguiar, H.G.M., Rangel, M.D., Azevedo, D.A., Aquino Neto, F.R., 2011. Comprehensive two-dimensional gas chromatography with time of flight mass spectrometry applied to biomarker analysis of oils from Colombia. Fuel 90, 2694–2699. https://doi.org/10.1016/j.fuel.2011.04.026.

da Silva, N., Cook, D., Lee, E.-K., 2017. Interactive Graphics for Visually Diagnosing Forest Classifiers in R. ArXiv170402502 Stat.

Smith, C.A., Maille, G.O., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., Siuzdak, G., 2005. METLIN: a metabolite mass spectral database. Ther. Drug Monit. 27, 747–751. https://doi.org/10.1097/01.ftd.0000179845.53213.39.

Špánik, I., Machyňáková, A., 2018. Recent applications of gas chromatography with high-resolution mass spectrometry. J. Sep. Sci. 41, 163–179. https://doi.org/10.1002/jssc.201701016.

Spicer, R.A., Salek, R.M., Moreno, P., Cañueto, D., Steinbeck, C., 2017. Navigating freely-available software tools for metabolomics analysis. Metabolomics 13. https://doi.org/10.1007/s11306-017-1242-7.

Stoll, D.R., Carr, P.W., 2017. Two-dimensional liquid chromatography: a state of the art tutorial. Anal. Chem. 89, 519–531. https://doi.org/10.1021/acs.analchem.6b03506.

Stubbins, A., Spencer, R.G.M., Chen, H., Hatcher, P.G., Mopper, K., Hernes, P.J., Mwamba, V.L., Mangangu, A.M., Wabakanghanzi, J.N., Six, J., 2010. Illuminated darkness: molecular signatures of Congo River dissolved organic matter and its photochemical alteration as revealed by ultrahigh precision mass spectrometry. Limnol. Oceanogr. 55, 1467–1477. https://doi.org/10.4319/lo.2010.55.4.1467.

Sumner, L.W., Amberg, A., Barrett, D., Beale, M.H., Beger, R., Daykin, C.A., Fan, T.W.-M., Fiehn, O., Goodacre, R., Griffin, J.L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A.N., Lindon, J.C., Marriott, P., Nicholls, A.W., Reily, M.D., Thaden, J.J., Viant, M.R., 2007. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics 3, 211–221. https://doi.org/10.1007/s11306-007-0082-2.

Swenson, T.L., Jenkins, S., Bowen, B.P., Northen, T.R., 2015. Untargeted soil metabolomics methods for analysis of extractable organic matter. Soil Biol. Biochem. 80, 189–198. https://doi.org/10.1016/j.soilbio.2014.10.007.

Swenson, T.L., Karaoz, U., Swenson, J.M., Bowen, B.P., Northen, T.R., 2018. Linking soil biology and chemistry in biological soil crust using isolate exometabolomics. Nat. Commun. 9. https://doi.org/10.1038/s41467-017-02356-9.

Szczepańska, N., Rutkowska, M., Owczarek, K., Płotka-Wasylka, J., Namieśnik, J., 2018. Main complications connected with detection, identification and determination of trace organic constituents in complex matrix samples. TrAC Trends Anal. Chem. 105, 173–184. https://doi.org/10.1016/j.trac.2018.05.005.

Tautenhahn, R., Cho, K., Uritboonthai, W., Zhu, Z., Patti, G.J., Siuzdak, G., 2012a. An accelerated workflow for untargeted metabolomics using the METLIN database. Nat. Biotechnol. 30, 826–828. https://doi.org/10.1038/nbt.2348.

Tautenhahn, R., Patti, G.J., Rinehart, D., Siuzdak, G., 2012b. XCMS online: a web-based platform to process untargeted metabolomic data. Anal. Chem. 84, 5035–5039. https://doi.org/10.1021/ac300698c.

Tfaily, M.M., Chu, R.K., Tolić, N., Roscioli, K.M., Anderton, C.R., Paša-Tolić, L., Robinson, E.W., Hess, N.J., 2015. Advanced solvent based methods for molecular characterization of soil organic matter by high-resolution mass spectrometry. Anal. Chem. 87, 5206–5215. https://doi.org/10.1021/acs.analchem.5b00116.

Tfaily, M.M., Chu, R.K., Toyoda, J., Tolić, N., Robinson, E.W., Paša-Tolić, L., Hess, N.J., 2017. Sequential extraction protocol for organic matter from soils and sediments using high resolution mass spectrometry. Anal. Chim. Acta 972, 54–61. https://doi.org/10.1016/j.aca.2017.03.031.

Thévenot, E.A., Roux, A., Xu, Y., Ezan, E., Junot, C., 2015. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for Univariate and OPLS statistical analyses. J. Proteome Res. 14, 3322–3335. https://doi.org/10.1021/acs.jproteome.5b00354.

Tranchida, P.Q., Aloisi, I., Giocastro, B., Mondello, L., 2018. Current state of comprehensive two-dimensional gas chromatography-mass spectrometry with focus on processes of ionization. TrAC Trends Anal. Chem. https://doi.org/10.1016/j.trac.2018.05.016.

Uclés, S., Lozano, A., Sosa, A., Parrilla Vázquez, P., Valverde, A., Fernández-Alba, A.R., 2017. Matrix interference evaluation employing GC and LC coupled to triple quadrupole tandem mass spectrometry. Talanta 174, 72–81. https://doi.org/10.1016/j.talanta.2017.05.068.

Van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J., 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7, 1–15. https://doi.org/10.1186/1471-2164-7-142.

Veenaas, C., Haglund, P., 2017. Methodology for non-target screening of sewage sludge using comprehensive two-dimensional gas chromatography coupled to high-resolution mass spectrometry. Anal. Bioanal. Chem. 409, 4867–4883. https://doi.org/10.1007/s00216-017-0429-0.

Verkh, Y., Rozman, M., Petrovic, M., 2018. Extraction and cleansing of data for a non-targeted analysis of high-resolution mass spectrometry data of wastewater. MethodsX 5, 395–402. https://doi.org/10.1016/j.mex.2018.04.008.

Vinaixa, M., Schymanski, E.L., Neumann, S., Navarro, M., Salek, R.M., Yanes, O., 2016. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: state of the field and future prospects. TrAC Trends Anal. Chem. 78, 23–35. https://doi.org/10.1016/j.trac.2015.09.005.

Walters, C.C., Wang, F.C., Higgins, M.B., Madincea, M.E., 2018. Universal biomarker analysis using GC × GC with dual FID and ToF-MS (EI/FI) detection. Org. Geochem. 115, 57–66. https://doi.org/10.1016/j.orggeochem.2017.10.003.

Wang, Y., Gu, M., 2010. The concept of spectral accuracy for MS. Anal. Chem. 82, 7055–7062. https://doi.org/10.1021/ac100888b.

Wang, Z., Stout, S.A., Fingas, M., 2006. Forensic fingerprinting of biomarkers for oil spill characterization and source identification. Environ. Forensic 7, 105–146. https://doi.org/10.1080/15275920600667104.

Wang, S., Li, J., Shi, X., Qiao, L., Lu, X., Xu, G., 2013. A novel stop-flow two-dimensional liquid chromatography–mass spectrometry method for lipid analysis. J. Chromatogr. A 1321, 65–72. https://doi.org/10.1016/j.chroma.2013.10.069.

Wang, J.-J., Pisani, O., Lin, L.H., Lun, O.O.Y., Bowden, R.D., Lajtha, K., Simpson, A.J., Simpson, M.J., 2017. Long-term litter manipulation alters soil organic matter turnover in a temperate deciduous forest. Sci. Total Environ. 607–608, 865–875. https://doi.org/10.1016/j.scitotenv.2017.07.063.

Ward, C.P., Cory, R.M., 2015. Chemical composition of dissolved organic matter draining permafrost soils. Geochim. Cosmochim. Acta 167, 63–79. https://doi.org/10.1016/j.gca.2015.07.001.

Warren, C.R., 2015. Comparison of methods for extraction of organic N monomers from soil microbial biomass. Soil Biol. Biochem. 81, 67–76. https://doi.org/10.1016/j.soilbio.2014.11.005.

Weber, R.J.M., Lawson, T.N., Salek, R.M., Ebbels, T.M.D., Glen, R.C., Goodacre, R., Griffin, J.L., Haug, K., Koulman, A., Moreno, P., Ralser, M., Steinbeck, C., Dunn, W.B., Viant, M.R., 2017. Computational tools and workflows in metabolomics: an international survey highlights the opportunity for harmonisation through galaxy. Metabolomics 13, 12. https://doi.org/10.1007/s11306-016-1147-x.

Wehrens, R., Hageman, J.A., van Eeuwijk, F., Kooke, R., Flood, P.J., Wijnker, E., Keurentjes, J.J.B., Lommen, A., van Eekelen, H.D.L.M., Hall, R.D., Mumm, R., de Vos, R.C.H., 2016. Improved batch correction in untargeted MS-based metabolomics. Metabolomics 12. https://doi.org/10.1007/s11306-016-1015-8.

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., Ni, Y., 2018. Missing value imputation approach for mass spectrometry-based metabolomics data. Sci. Rep. 8. https://doi.org/10.1038/s41598-017-19120-0.

Westerhuis, J.A., Hoefsloot, H.C.J., Smit, S., Vis, D.J., Smilde, A.K., van Velzen, E.J.J., van Duijnhoven, J.P.M., van Dorsten, F.A., 2008. Assessment of PLSDA cross validation. Metabolomics 4, 81–89. https://doi.org/10.1007/s11306-007-0099-6.

Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., Scalbert, A., 2013. HMDB 3.0—the human metabolome database in 2013. Nucleic Acids Res. 41, D801–D807. https://doi.org/10.1093/nar/gks1065.

Wrona, M., Mauriala, T., Bateman, K.P., Mortishire-Smith, R.J., O'Connor, D., 2005. 'All-in-One' analysis for metabolite identification using liquid chromatography/hybrid quadrupole time-of-flight mass spectrometry with collision energy switching. Rapid Commun. Mass Spectrom. 19, 2597–2602. https://doi.org/10.1002/rcm.2101.

Wu, Y., Li, L., 2016. Sample normalization methods in quantitative metabolomics. J. Chromatogr. A 1430, 80–95. https://doi.org/10.1016/j.chroma.2015.12.007.

Xi, B., Gu, H., Baniasadi, H., Raftery, D., 2014. Statistical analysis and modeling of mass spectrometry-based metabolomics data. Methods Mol. Biol. 1198, 333–353. https://doi.org/10.1007/978-1-4939-1258-2_22.

Xia, J., Wishart, D.S., 2011. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. Nat. Protoc. 6, 743–760. https://doi.org/10.1038/nprot.2011.319.

Xia, J., Wishart, D.S., 2016. Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. Curr. Protoc. Bioinformatics 55, 14.10.1–14.10.91. https://doi.org/10.1002/cpbi.11.

Xu, Y., Zomer, S., Brereton, R.G., 2006. Support vector machines: a recent method for classification in chemometrics. Crit. Rev. Anal. Chem. 36, 177–188. https://doi.org/10.1080/10408340600969486.

Yadav, S.P., 2007. The wholeness in suffix -omics, -omes, and the word Om. J. Biomol. Tech. 18, 277.

Yang, C., Wang, Z., Yang, Z., Hollebone, B., Brown, C.E., Landriault, M., Fieldhouse, B., 2011. Chemical fingerprints of Alberta oil sands and related petroleum products. Environ. Forensic 12, 173–188. https://doi.org/10.1080/15275922.2011.574312.

Yang, J., Zhao, X., Lu, X., Lin, X., Xu, G., 2015. A data preprocessing strategy for metabolomics to reduce the mask effect in data analysis. Front. Mol. Biosci. 2. https://doi.org/10.3389/fmolb.2015.00004.

Yao, C.-H., Liu, G.-Y., Yang, K., Gross, R., Patti, G., 2016. Inaccurate quantitation of palmitate in metabolomics and isotope tracer studies due to plastics. Metabolomics 12, 1–7. https://doi.org/10.1007/s11306-016-1081-y.

Yi, L., Dong, N., Yun, Y., Deng, B., Ren, D., Liu, S., Liang, Y., 2016. Chemometric methods in data processing of mass spectrometry-based metabolomics: a review. Anal. Chim. Acta 914, 17–34. https://doi.org/10.1016/j.aca.2016.02.001.

Zhang, W., Zhu, S., He, S., Wang, Y., 2015. Screening of oil sources by using comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry and multivariate statistical analysis. J. Chromatogr. A 1380, 162–170. https://doi.org/10.1016/j.chroma.2014.12.068.

Zhang, X., Quinn, K., Cruickshank-Quinn, C., Reisdorph, R., Reisdorph, N., 2018. The application of ion mobility mass spectrometry to metabolomics. Curr. Opin. Chem. Biol. 42, 60–66. https://doi.org/10.1016/j.cbpa.2017.11.001 Omics.

Zhao, Y.-Y., Wu, S.-P., Liu, S., Zhang, Y., Lin, R.-C., 2014. Ultra-performance liquid chromatography–mass spectrometry as a sensitive and powerful technology in lipidomic applications. Chem. Biol. Interact. 220, 181–192.

Zheng, X., Aly, N.A., Zhou, Y., Dupuis, K.T., Bilbao, A., Paurus, V.L., Orton, D.J., Wilson, R., Payne, S.H., Smith, R.D., Baker, E.S., 2017. A structural examination and collision cross section database for over 500 metabolites and xenobiotics using drift tube ion mobility spectrometry. Chem. Sci. 8, 7724–7736. https://doi.org/10.1039/C7SC03464D.